# Statistical Methods and Modeling in Cancer Etiology and Early Detection

**by**

**Lu Li**

**A dissertation submitted to The Johns Hopkins University**

**in conformity with the requirements for the degree of**

**Doctor of Philosophy**

**Baltimore, Maryland**

**November, 2018**

# Abstract

Notwithstanding the many advances made in cancer research over the last several decades, there are still many fundamental questions that need to be addressed. During my PhD, I have been working on developing statistical methods to analyze liquid biopsy data for cancer early detection and using mathematical modeling to better understand cancer etiology.

Traditionally, cancer-causing mutations have been thought to have two major sources: inherited (H) and environmental (E) risk factors. However, these are not enough to explain the extreme variation in cancer incidence across different tissues. Recently, the random mutations occurring during normal cell replications (R mutations) were recognized as a third and important source of cancers. In the first part of this dissertation, we proposed a novel approach to estimate the proportions of these three sources of mutations using cancer genome sequencing and epidemiological data. Our method suggests that R mutations are responsible for two-thirds of the mutations in human cancers or at least not less than 40% of them.

At the same time, while the role of driver mutations and other genomic alterations in cancer causation is well recognized, they may not be sufficient for cancer to occur. Other factors like, for example, the immune system

and the microenvironment, also have their impacts. It's not known, as of now, how large is the contribution of mutations compared to all these other factors, which we collectively define as K factors. Therefore, the second part of this dissertation is trying to address this question. We develop a method to estimate how much of the observed increase in cancer risk due to a E or H factor can be explained by the increase in mutation rate caused by that factor, thus providing an assessment of the role of mutations during tumorigenesis. Genome sequencing and epidemiological data are used to perform the analysis. Our results show that the higher mutation rate is able to explain the majority of the increase in cancer risk due to smoking and microsatellite instability (MSI), a moderate fraction in Hepatitis C (HCV) infections, but almost nothing when considering obesity.

Overall, the above results indicate that a relevant amount of cancers may not be preventable due to the unavoidable R mutations. Thus, in addition to finding new therapies, it seems critical to be able to detect cancers earlier in order to reduce cancer deaths. In the third part of this dissertation, we develop statistical methods to analyze genome sequencing data for the early detection of cancer. Combining mutational information with protein assays, the median sensitivity of our method is 70% in eight cancer types, including ovary, liver, stomach, pancreas, esophagus, colorectum, lung, and breast, with 99% specificity. The sensitivities range from 69% to 98% for the detection of five cancer types (ovary, liver, stomach, pancreas, and esophagus) for which there are no screening tests available for average-risk individuals.

# Thesis Committee

## Primary Readers

John Groopman (Chair)
Professor
Department of Environmental Health and Engineering
Johns Hopkins Bloomberg School of Public Health

Cristian Tomasetti (Primary Advisor)
Associate Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Hongkai Ji
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Nickolas Papadopoulos
Professor
Department of Oncology
Johns Hopkins School of Medicine

## Alternate Readers

Brian Caffo
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Corinne E. Joshu
Associate Professor
Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health

# Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Cristian Tomasetti, for his endless and unconditional support throughout these years, not only in academic research, but also in personal development. Dr. Tomasetti has been always a great encouragement to me through difficulties, and a great inspiration, both intellectually as a researcher and more generally as a mentor. He showed up on my first presentation to make me feel less nervous. He shared his own experiences to help me make decision on my future career. More importantly, he teaches me how to approach a seemingly obvious problem from a novel perspective and to work independently. He encourages me to explore more in the field of translational research and provides me with lots of great opportunities to work on interesting projects. He is the most supportive advisor I could ever imagine. I'm proud of being his first PhD student.

I also greatly appreciate the help, friendship and support from my collaborators, Drs. Vogelstein, Kinzler, Papadopoulos and their lab members from the Ludwig Center at Johns Hopkins. I'm thrilled to have the honor of working with the pioneers in the field of cancer genomics. I've learned not only the knowledge, but also how to be a good researcher. I would also like to mention the lab members Yuxuan Wang, Joshua Cohen and Simeon Springer,

who have provided huge support during our collaborations.

Thanks to all the thesis and oral committee members Drs. John Groopman, Brian Caffo, Elizabeth Platz, Ken Kinzler, Hongkai Ji, Nick Papadoupoulous, Ciprian Crainiceanu and Corinne E. Joshu, for their precious time and constructive advice.

Special thanks to my colleagues in the Biostatistics department: Yu Du, Junrui Di, Haoyu Zhang, Youjin Lee, Jiawei Bai, Yuxin Zhu, Detian Deng and many others that I cannot list here, as well as members from Dr. Tomasetti's lab: Bahman Afsari, Ludmila Danilova, Yifan Zhang, Kamel Lahouel. The inspiring conversations and valuable peer suggestions are tremendously helpful.

Finally, I want to thank my parents, who always stand behind me providing everything I need. I wouldn't survive without their love and support. It's a pity that I have spent little time with them in the past 10 years, but my heart stays with them all the time.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Cancer mortality, etiology, and prevention

It is estimated that the number of new cancer cases for 2018 in the US is over 1.7 million (*SEER*). With an aging population, this number is likely to increase in the future. 609,640 Americans are expected to die of cancer in 2018, which makes cancer the second most common causes of death in the US, only exceeded by heart disease (*Cancer Facts and Figures 2018*). And when considering the last fifty years, the decrease in cancer mortality rates among developed countries is much smaller than the decrease in mortality rates observed for heart disease (Wang et al., 2016). From a public health perspective, this then presents the challenge of having better strategies on cancer prevention, which in turn requires a better knowledge of cancer etiology. Environmental (E) and inherited factors (H) are by far the most well-accepted cancer causes. For example, an estimate of 80% to 90% of the lung cancers in the US are linked to cigarette smoking (*Centers for Disease Control and Prevention*). The role of hereditary factors have been demonstrated from both twin studies (Mucci

et al., 2016) and the identification of the genes responsible for cancer predisposition syndromes (Vogelstein et al., 2013). Recently, a third source - mutations due to the random mistakes made during normal DNA replication (R) - has been proposed to explain the variation in cancer risk among different tissues (Tomasetti and Vogelstein, 2015), and overall R mutations seem responsible for a large fraction of the mutations found in human cancers(Tomasetti, Li, and Vogelstein, 2017).

Prevention of any disease can be classified into primary prevention and secondary prevention (Song et al., 2018). Primary prevention is the intervention before any disease happens, by using vaccines, avoiding risky behaviors and substances known to be related to a disease, while secondary prevention is the detection and intervention at the early stage of the disease. Primary prevention is undoubtedly the best way to reduce cancer deaths, but is not currently available for all cancer types. For cancers without effective primary prevention strategies, early detection appears to be the key to reduce cancer deaths. With the advancement of sequencing technologies, liquid biopsies are becoming an important and promising research direction, given the benefits of being non-invasive, relatively quick, and easily repeatable tests. An example of a recent blood test developed for cancer early detection is (Cohen et al., 2018). The new type of data generated by liquid biopsies require the development of new methods for their analysis.

## 1.2 Structure of the dissertation

In this dissertation, we apply modeling techniques to address some of the important questions on cancer etiology and develop statistical methods for liquid biopsy data in cancer early detection. The rest of the thesis is organized as follows.

In Chapter 2, we introduce a novel approach to quantify the fractions of mutations due to environmental factors, inherited factors and random mutations. We used epidemiology data and genome sequencing results to evaluate the hypothesis that R mutations play a major role in cancer. The result explicitly and quantitatively address the difference between cancer etiology and cancer preventability.

In Chapter 3, we propose a method to assess the role of mutations in cancer causation. It is known that mutations represent a necessary ingredient for cancer to occur. But it may not be sufficient. This part of the work is trying to find how large is the contribution of mutations when compared to other important factors like, for example, the immune system and the microenvironment.

In Chapter 4, we illustrate a method for analyzing the data from a blood test for the early detection of cancers of ovary, liver, stomach, pancreas, esophagus, colorectum, lung and breast. This method provides of a median sensitivity of 70% among the eight cancer types, at the specificity level of 99%. The sensitivities ranged from 69% to 98% for cancers (ovary, liver, stomach, pancreas, and esophagus) for which no screening tests are currently available for

average-risk individuals.

Discussions and future work are provided in Chapter 5.

# References


*SEER*. https://seer.cancer.gov/statfacts/html/all.html.

*Cancer Facts and Figures 2018*. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2018.html.

Wang, Haidong, Mohsen Naghavi, Christine Allen, Ryan M Barber, Zulfiqar A Bhutta, Austin Carter, Daniel C Casey, Fiona J Charlson, Alan Zian Chen, Matthew M Coates, et al. (2016). "Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015". In: *The lancet* 388.10053, pp. 1459–1544.

*Centers for Disease Control and Prevention*. https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm.

Mucci, Lorelei A, Jacob B Hjelmborg, Jennifer R Harris, Kamila Czene, David J Havelick, Thomas Scheike, Rebecca E Graff, Klaus Holst, Sören Möller, Robert H Unger, et al. (2016). "Familial risk and heritability of cancer among twins in Nordic countries". In: *Jama* 315.1, pp. 68–76.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler (2013). "Cancer genome landscapes". In: *science* 339.6127, pp. 1546–1558.

Tomasetti, Cristian and Bert Vogelstein (2015). "Variation in cancer risk among tissues can be explained by the number of stem cell divisions". In: *Science* 347.6217, pp. 78–81.

Tomasetti, Cristian, Lu Li, and Bert Vogelstein (2017). "Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention". In: *Science* 355.6331, pp. 1330–1334.

Song, Mingyang, Bert Vogelstein, Edward L Giovannucci, Walter C Willett, and Cristian Tomasetti (2018). "Cancer prevention: Molecular and epidemiologic consensus". In: *Science* 361.6409, pp. 1317–1318.

Cohen, Joshua D, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A Javed, Fay Wong, Austin Mattox, et al. (2018). "Detection and localization of surgically resectable cancers with a multi-analyte blood test". In: *Science*, eaar3247.

# Chapter 2

# Estimation of proportions of mutations using EHR model

## 2.1 Introduction

It is now widely accepted that cancer is the result of the gradual accumulation of driver gene mutations that successively increase cell proliferation (Garraway and Lander, 2013; Stratton, Campbell, and Futreal, 2009; Vogelstein et al., 2013). But what causes these mutations? The role of environmental factors (E) in cancer development has long been evident from epidemiological studies, and this has fundamental implications for primary prevention. The role of heredity (H) has been conclusively demonstrated from both twin studies (Mucci et al., 2016) and the identification of the genes responsible for cancer predisposition syndromes (Vogelstein et al., 2013; Stadler et al., 2010). A recently study hypothesized that a third source - mutations due to the random mistakes made during normal DNA replication (R) - can explain why cancers occur much more commonly in some tissues than others (Tomasetti and Vogelstein, 2015b). This hypothesis was based on the observation that,

in the United States, the lifetime risks of cancer among 25 different tissues were strongly correlated with the total number of divisions of the normal stem cells in those tissues (Tomasetti and Vogelstein, 2015b; Tomasetti and Vogelstein, 2015a). It has been extensively documented that approximately three mutations occur every time a normal human stem cell divides (Lynch, 2010; Tomasetti, Vogelstein, and Parmigiani, 2013). Therefore, the authors inferred that the root causes of the correlation between stem cell divisions and cancer incidence were the driver gene mutations that randomly result from these divisions. Recent evidence from mouse models supports the notion that the number of normal cell divisions dictates cancer risk in many organs (Zhu et al., 2016).

However, this analysis was confined to explaining the relative risk of cancer among tissues rather than the contribution of each of the three potential sources of mutations (E, H, and R) to any single cancer type or cancer case (Tomasetti and Vogelstein, 2015b). Determination of the contributions of E, H, and R to a cancer type or cancer case is challenging. In some patients, the contribution of H or R factors might be high enough to cause all the mutations required for that patient's cancer, whereas in others, some of the mutations could be due to H, some to R, and the remainder to E. Here we perform a critical evaluation of the hypothesis that R mutations play a major role in cancer. Our evaluation is predicated on the expectation that the number of endogenous mutations (R) resulting from stem cell divisions in a tissue, unlike those caused by environmental exposures, would be similarly distributed at a given age across human populations. Though the number of stem cell

divisions may vary with genetic constitution (e.g., taller individuals may have more stem cells), these divisions are programmed into our species' developmental patterns. In contrast, deleterious environmental and inherited factors, either of which can directly increase the mutation rate or the number of stem cell divisions, vary widely among individuals and across populations. To perform the evaluation, we developed a novel approach to determine what fractions of cancer-causing mutations result from E, H, or R. These fractions have not been estimated for any cancer type previously.

## 2.2 Method

### 2.2.1 Single Environmental Factor

We define $T$ as the total number of (clonal) somatic mutations present in the tumor cell that acquired the last of the driver gene mutations required to yield a given cancer, i.e. the founding cell of the final clonal expansion. First, consider cancers in patients not affected by any deleterious environmental (E) or inherited (H) factor. Given our definition of replicative (R) mutations, the number of somatic mutations that occur in a tissue independently of the deleterious effects of environmental (E) or inherited (H) factors are assigned to R. It follows that in those cancers from patients not exposed to any E and H factors, all the required driver mutations, as well as all passenger mutations, occurring before and during the tumor's clonal expansions are due to R. The total number of somatic mutations in the first cancer cell (the founder cell) of

those patients is

$$T_R = \sum_{d=1}^{M} m_d$$

where $M$ is the total number of divisions of the cell until it acquired the last driver gene mutation, and $m_d$ is the number of somatic mutations occurring at $d^{th}$ cell division, with $d = 1, ..., M$. Note that both $M$ and all $m_d$ are random variables, and that the resulting random variable $T_R > 0$ because at least one driver gene mutation is required to develop a cancer. By considering a population of cancer patients only affected by R, we obtain a distribution for the total number of somatic mutations, $T_R$, in the unexposed population.

Next, consider the case where only one environmental (E) factor and no inherited (H) factors affect a cancer's incidence. For patients exposed to this harmful environmental factor E at the same level, the total number of somatic mutations in the founding cancer cell, $T_E$, is

$$T_E = \sum_{d=1}^{M'} m_d^{'}$$

with $M' \geq_{st} M$, or $m' \geq_{st} m$, or both, where $\geq_{st}$ stands for stochastically greater and $T_E$ >0.

Intuitively, E is expected to either increase the number of somatic mutations that occur in each cell division (e.g., a mutagen that acts directly on DNA) or to increase the rate of cell division itself (e.g., smoking can inflame the lungs, causing higher cell turnover), or both, thus increasing the rate at which the cells acquire somatic mutations. It does not matter which one of the two occurred or whether both occurred. What matters is that it occurred, as measured through sequencing. The total number of somatic mutations, which

are almost all passenger mutations, are akin to a clock that measures both the number of divisions and the mutation probabilities: doubling the number of cell divisions yields on average two times more mutations after a given period, just as doubling the mutation rate does over that time period, all other things being equal. Similarly, a doubling of the division rate (or a doubling of the probability of mutation per cell division per base), if applied only to half of the time interval before the founder cell is born, yields $\frac{1}{2} * 1 + \frac{1}{2} * 2 = 1.5$ times more mutations on average, all other things being equal. Note that even the effect of a single event, occurring in a very short time interval (e.g. seconds), such as a one-time exposure to a strong radiation source increasing $m_d$ by many fold, will be appropriately recorded in this sum. Basically, we do not know the specific values for $M$ and each $m_d$, but from sequencing data we can observe the output of a sum, $T_E$, which summarizes the mutational life history of the cancer cell's lineage.

When comparing $T_E$ and $T_R$ in cancer patients diagnosed at the same age $a$, $T_E(a) \geq_{st} T_R(a)$ because the incidence of that cancer in the exposed group is greater than in the unexposed group. Consider now what happens when we sample two cancer patients $P_1$ and $P_2$ diagnosed at the same age, one who was exposed to E ($P_1$) and one who was not ($P_2$), where $t_E$ is the specific realization of $T_E(a)$ in $P_1$ and $t_R$ is the realization of $T_R(a)$ in $P_2$, with $t_E(a) \geq t_R(a)$. We can split the mutation load in the exposed patient $P_1$ into two proportions

$$1 = \frac{t_R(a)}{t_E(a)} + \frac{t_E(a) - t_R(a)}{t_E(a)}$$

where the first term in the sum is the proportion of mutations that patient $P_1$

would have had even if unexposed to E (i.e. attributable to R), assuming $P_1$ would have had in that case the same total mutational load of $P_2$, while the second term in the sum is the proportion of extra mutations that $P_1$ acquired due to exposure to E. Obviously we cannot assume that $P_1$ would have had exactly the same mutational load of $P_2$ if $P_1$ was unexposed, because both $T_E(a)$ and $T_R(a)$ are random variables rather than single fixed values, and in fact it is possible that $T_E(a) < T_R(a)$ in some pairs of patients. However, given the distribution of $T_R(a)$ among all unexposed patients, it is true that, for patient $P_1$ with $T_E = t_E$,

$$1 = \frac{T_R(a)}{t_E(a)} + \frac{t_E(a) - T_R(a)}{t_E(a)}$$

and the distribution of the random variable $\min(\frac{T_R(a)}{t_E(a)}, 1)$ is the distribution of the proportion of mutations attributable to R in patient $P_1$. Analogously, $1 - \min(\frac{T_R(a)}{t_E(a)}, 1)$ is the distribution of the proportion of mutations attributable to E in patient $P_1$. As $T_E(a) \geq_{st} T_R(a)$, the average proportion of mutations attributable to R and to E at the population level can be estimated by taking the expected value of the above expression (in Equation 2.1 below, "E" denotes the mathematical expectation operator and has nothing to do with environmental factors):

$$1 = \mathbf{E}(\frac{T_R(a)}{T_E(a)}) + \mathbf{E}(\frac{T_E(a) - T_R(a)}{T_E(a)}) \tag{2.1}$$

Therefore, for the population of patients of age $a$ that were exposed to environmental factor E, we can on average attribute the first term in the sum to R, i.e. assign the average proportion $E(\frac{T_R(a)}{T_E(a)})$ of the overall mutation load to R, and attribute the second average proportion to environmental factor

E. Importantly, these same proportions can be used to attribute the average proportions of driver gene mutations (rather than total mutations) to R and E at the population level.

In practice, we will use the median instead of the mean for robustness - the distribution of the total number of somatic mutations in tumors typically has a large right tail (e.g. due to patients with defective mismatch repair mechanisms). And we will control for age by dividing the specific $T$ of patients by their ages $a$. Given $n$ cancer patients exposed to E and $m$ cancer patients not exposed to E, we define MR as the median value in the population of the ratio,

$$MR = median(\frac{T_{E_i}/a_i}{T_{R_i}/a_j}), i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}$$

with its 95% confidence interval, as estimated via bootstrap. $MR > 1$ implies that a larger average number of mutations had accumulated in the founder cells of the cancers of the patients exposed to E than in unexposed patients after controlling for age. In other words, cancers in patients exposed to E had a larger number of somatic mutations per year. Then, similarly to Equation 2.1,

$$1 = \frac{1}{MR} + (1 - \frac{1}{MR})$$

As before, it follows that in those exposed to the E factor, $\frac{1}{MR}$ of the mutation load can be attributed on average to R, and $\frac{MR-1}{MR}$ can be attributed to the extra effects of E. The same is true for the proportion of driver gene mutations (rather than total mutational load) attributable to R and E factors. Thus, we define MEPAR, the mutational extrinsic proportion of attributable risk caused

by the extrinsic factor E in a given cancer type, as

$$MEPAR = \frac{MR - 1}{MR}$$

MEPAR represents the median proportion of driver genes mutations that are attributable to the extrinsic factor E in patients that are exposed to E.

For a given cancer type, we further define $PPE$, the proportion of cancer patients exposed to environmental factor E, as

$$PPE = \frac{P_e \cdot RR}{1 + P_e \cdot (RR - 1)}$$

where $P_e$ is the prevalence of E in the general population, $RR$ is the relative risk of people getting a given cancer due to exposure of E comparing to people unexposed to E. Thus, to estimate $P_E$, the fraction of mutations attributed to E in the cancer population, we have

$$P_E = PPE \cdot MEPAR = PPE \cdot \frac{MR - 1}{MR} \tag{2.2}$$

and the proportion $P_R$ of the driver mutations required to develop this type of cancer and attributable to R, is given by

$$P_R = 1 - MEPAR \cdot PPE = \frac{1}{MR} \cdot PPE + 1 \cdot (1 - PPE) \tag{2.3}$$

The first term in the sum on the right side of Equation 2.3 is the proportion of driver gene mutations attributable to R in patients exposed to E, and the second term in the sum is the proportion of the driver gene mutations caused by R in patients not exposed to E (which is always 100%). Note that this equation does not require that every cancer patient has the same number

of required driver mutations as long as the distribution of the number of required driver mutations is the same in patients with a given cancer type, independently of E. Also, the same method can in principle be applied to multiple quantiles of the distribution of the ratio rather than only to the median $MR$. For example, if ten equidistant quantiles are used, each of those quantile values will have a 0.1 weight in estimating the ratio. Importantly, for the examples provided below, using the median $MR$ is essentially the same or more conservative than using multiple quantiles.

If we let $P_E$ be the proportion of driver gene mutations due to the E factor, i.e. $P_E = 1 - P_R$, then the proportion of attributable risk (PAR, also known as PAF), is

$$PAR = \frac{P_e \cdot (RR - 1)}{1 + P_e \cdot (RR - 1)} = \frac{RR - 1}{RR} \cdot PPE$$

It follows from Equation 2.2 that

$$PAR = \frac{RR - 1}{RR} \cdot \frac{MR}{MR - 1} \cdot P_E$$

This expression indicates the relationship between the proportion of cancers cases that are preventable, PAR, and the proportion of driver gene mutations that are attributable to the E factor, i.e. the relationship between cancer preventability and cancer etiology. Importantly, PAR$> P_E$, given that RR $>$ MR.

### 2.2.2 Generalization to Multiple Factors

Until now, we have been looking at the simplest case where only one environmental factor is present for a given cancer type. Now, we generalize

the above results to a more practical scenario, where multiple environmental factors could affect the incidence of a specific cancer.

Let $k, j, e$ be indexes respectively for the cancer type, age group and E factor, among the possible E factors known to affect cancer type $k$. Let $w_j$ be the proportion of the general population for age group $j$, $I_{k,j}$ be the incidence of cancer type $k$ among population for age group $j$. Let $i_e$ be the exposure level of factor E (there might be different levels of one factor, e.g. exercise level). Consider for each cancer type $k$ and for each age group $j$, a partition with index $\delta_{k,j}$ of the cancer patient population in that age group, where each element of the partition represents patients with a unique combination of exposure levels across all environmental factors that have an effect on cancer type $k$. The total number of partitions is given by the numbers of environmental factors associated with cancer $k$ and the numbers of levels for each related factors. For example, if there are three factors known to have effects on cancer type $k$, and these three factors each has 3,4,5 different exposure levels respectively, then the total number of partitions is $3 \cdot 4 \cdot 5 = 60$, i.e. $\delta_{k,j}$ will assume values from 1 to 60.

Define $P_{\delta_{k,j}}$ as the proportion of patient with cancer type $k$ within age group $j$ that is in sub group $\delta_{k,j}$,

$$P_{\delta_{k,j}} = \prod_e PPE_{k,j,i_e} = \prod_e \frac{P_{e_{j,i_e}} RR_{k,i_e}}{1 + \sum_i (P_{e_{j,i_e}} (RR_{k,i_e} - 1))}$$

where $P_{e_{j,i_e}}$, according to the definition of $PPE$, is the prevalence in the general population, in age group $j$ with E factor $e$ at exposure level $i_e$, and $RR_{k,i_e}$ is the relative risk of cancer $k$ for those exposed to factor $e$ at exposure level $i_e$.

Define $MR_{k,i_e}$ as the median ratio of the number of mutations, adjusted for age, of those exposed to factor $e$ at exposure level $i_e$, and those unexposed, in cancer $k$. To be conservative, we assume that different factors have multiplicative effects on cancer risk, then the proportion of drive gene mutations attributable to $e$ in cancer $k$ can be written as

$$\begin{aligned} P_{E_k} &= \sum_j \left( \frac{w_j I_{k,j}}{\sum_j (w_j I_{k,j})} P_{E_{k,j}} \right) \\ &= \sum_j \left( \frac{w_j I_{k,j}}{\sum_j (w_j I_{k,j})} \sum_{\delta_{k,j}} \left( 1 - \frac{1}{1 + \sum_e (MR_{k,i_e} - 1)} \right) P_{\delta_{k,j}} \right) \end{aligned} \tag{2.4}$$

Hence, we can get the estimate of overall proportion of driver gene mutations attributable to E across all cancer types, $P_E$, where each cancer type $k$ is weighted by its incidence relative to the overall cancer incidence

$$\begin{aligned} P_E &= \sum_j \left( \frac{w_j \sum_k I_{k,j}}{\sum_j (w_j I_{k,j})} \sum_k \left( \frac{I_{k,j}}{\sum_k I_{k,j}} P_{E_{k,j}} \right) \right) \\ &= \sum_j \left( \frac{w_j}{\sum_j (w_j \sum_k I_{k,j})} \sum_k \left( I_{k,j} P_{E_{k,j}} \right) \right) \end{aligned} \tag{2.5}$$

Having estimated $P_E$, the last step is to obtain $P_H$ (see Section 2.2.4 for the method). It then follows that the residual proportion of driver gene mutations can be attributed to $P_R$

$$P_R = 1 - (P_E + P_H) \tag{2.6}$$

Equations 2.4, 2.5 and 2.6 can be used to get the gender-specific estimates of $P_E$, $P_H$, and $P_R$. It is also possible to obtain the estimates combining both

sexes by weighting on the gender $s$ accordingly:

$$\begin{aligned} P_{E_{k,s}} &= \sum_j \left( \frac{w_{j,s} I_{k,j,s}}{\sum_j (w_{j,s} I_{k,j,s})} P_{E_{k,j,s}} \right) \\ &= \sum_j \left( \frac{w_{j,s} I_{k,j,s}}{\sum_j (w_{j,s} I_{k,j,s})} \sum_{\delta_{k,j,s}} \left( 1 - \frac{1}{1 + \sum_e (MR_{k,i_e,s} - 1)} \right) P_{\delta_{k,j,s}} \right) \end{aligned} \tag{2.7}$$

and

$$P_E = \sum_j \sum_s \left( \frac{w_{j,s}}{\sum_j \sum_s (w_{j,s} \sum_k I_{k,j,s})} \sum_k \left( I_{k,j,s} P_{E_{k,j,s}} \right) \right) \tag{2.8}$$

### 2.2.3 Determination of MR when cancer genome sequencing data is not available

Previously, we discuss how to get the estimate of MR when sequencing data is available. However, genome-wide sequencing data of cancers from patients exposed and unexposed to specific environmental factors is generally not available, so there is no experimentally derived estimate of MR. In this section, we describe the method that conservatively estimates MR for any environmental factor that is based simply on epidemiological data about the relative risk (RR) conferred by that factor. This approach is applicable to the very large number of cancer types that have been investigated extensively through sophisticated epidemiologic techniques (e.g., those summarized in Cancer Research UK, 2018b).

If $x$ is the increase in mutation rate resulting from exposure to an environmental factor, then a mathematical bound for $x$ is provided by $x \leq RR^{\frac{1}{2}}$ when two driver gene mutations are required for cancer formation (Tomasetti et al., 2015). Intuitively, a $x$-fold increase in the mutation rate would result in an $x^2$

fold increase in cancer incidence because of the well-known exponential rather than linear relationship between cancer incidence and age. The inequality reflects the fact that there may be other deleterious effects (e.g. epigenetic) induced by that environmental exposure that also increase RR by some factor c, either additively (i.e. $RR = c + x^2$ with $c > 0$) or multiplicatively (i.e. $RR = c \cdot x^2$, with $c > 1$). We estimated MRs under the assumption that there were no other deleterious effects; this assumption is conservative because deleterious effects would lower the estimated MRs. Similarly, when three mutations are required for cancer formation, we can assume $x \leq RR^{0.4}$. An exponent of 0.4 is used in this case by assuming that each successive driver gene mutation doubled the proliferative fitness advantage, as explained in (Tomasetti et al., 2015; Durrett and Moseley, 2010). For most common cancers, the minimum number of driver gene mutations required for cancer development is likely to be 3 (Tomasetti et al., 2015), and this minimum is strongly supported by genome-wide sequencing studies on tens of thousands of cancers (Garraway and Lander, 2013; Stratton, Campbell, and Futreal, 2009; Vogelstein et al., 2013).

When applied to lung adenocarcinomas and pancreatic ductal adenocarcinomas, the increase in mutation rates caused by smoking can then calculated to be $18^{0.4} = 3.2$ in lung adenocarcinomas, and $2^{0.4} = 1.32$ in pancreatic ductal adenocarcinomas. Though based exclusively on epidemiological data, these numbers are remarkably close to the estimates of $x$ obtained experimentally from genome-wide sequencing: 3.23 and 1.16 for lung adenocarcinomas and pancreatic ductal adenocarcinomas, respectively, providing experimental

validation of our heuristic.

### 2.2.4 Determining Oncogenic Viruses and Heredity Effects

Assume that a given cancer type requires $n$ driver mutations to occur and that those with a given hereditary predisposition to that cancer already have one of those required driver gene mutations. Then, $\frac{1}{n}$ of the required driver mutations would be caused by this H factor. To be conservative, we attributed $\frac{1}{2}$ of the total driver gene mutations to H in all these cases, even in solid cancers, in which a minimum of three driver gene mutations is likely to be necessary for cancer formation. Similarly, in those exposed to a virus, such as HPV 16 in patients with cervical cancers, we attribute $\frac{2}{3}$ of the required driver mutations to that factor. The effects of these oncogenic viruses can be considered to be mutations (i.e., changes in the DNA) because they are associated with the insertion of viral oncogenes into the nucleus. The reason for attributing $\frac{2}{3}$ to such viral infections is that the protein products of oncogenic viruses like HPV inactivate two driver gene pathways (TP53 and Rb; McLaughlin-Drubin, Meyers, and Munger, 2012). In other virally-associated cancers where the molecular pathogenesis is not as clear, such as in liver cancers associated with HBV infection, it is conservative to assume that $\frac{2}{3}$ of the driver gene "mutations" are contributed by the virus.

## 2.3 Data

The epidemiological information (prevalence $P_e$, relative risk $RR$, and incidence $I$) for E factors known to affect 32 cancer types was obtained from the

UK Cancer Research database (Cancer Research UK, 2018b; Cancer Research UK, 2018c; Parkin, 2011). Whenever the prevalence was not indicated for a given age interval (typically the youngest group), we conservatively attributed to that environmental factor the prevalence of the adjacent age group. Kaposi sarcoma was not included because incidence rates were not provided (an extremely rare cancer). Given that UK epidemiological data were used, we also used the weights $w$ of the general U.K. population (all individuals, not just cancer patients) for each age group and for each gender (UK Office for National Statistics, 2015). To be conservative, environmental factors for which epidemiological information was not fully available (specifically: second-hand smoking, radiation, lack of breastfeeding, and occupational exposures) were assumed to have the same, large effect on the mutation rate (i.e. same MR) that smoking has on lung cancer, with PPE approximated by their PAR. Due to the lack of data on the prevalence of various salt consumption levels, we set the average consumption level of salt with $PPE = 1$ and compared it to the recommended level of 6g per day to obtain MR. After obtaining estimates for the PPEs and MRs for all known risk factors, Equations 2.4 and 2.5 were used to obtain sex-specific estimates of $P_E$, for each cancer type. Similarly, Equations 2.7 and 2.8 above were used to obtain estimates of $P_E$ when combining the two sexes. Two thirds of the required driver gene mutations were attributed to viral infections in cancers associated with viruses, with PPE approximated by PAR, and added to the overall $P_E$.

Information on the proportion of cancer cases attributable to inherited factors (H) was obtained from the UK Cancer Research database (Cancer

Research UK, 2018a), and if no estimate was available, 1% was assumed. It is estimated that, overall, inherited mutations play a role in 5-10% of cancer cases (National Cancer Institute, 2018). Thus, 10% of cancer cases were overall conservatively attributed to H (i.e. PAR = 0.10) when deriving the estimates of $P_H$ combined over all cancer types. As described above, we conservatively attributed half of the driver gene mutations to H (i.e. $P_H = 0.5$) in every cancer case attributed to H.

Having estimates for $P_E$ and $P_H$, Equation 2.6 was then used to estimate $P_R$, in each cancer type. We assumed that the number of driver genes required for cancer development was three in 25 of the 32 cancer types. The seven exceptions were cancers of the bone, testes, myelomas, retinoblastomas, leukemias, Hodgkin and non-Hodgkin lymphomas, in which we assumed that two driver gene mutations were required for cancer. This assumption is based on molecular genetic and epidemiological data indicating that fewer driver gene mutations are required for the development of "liquid tumors" and pediatric cancers than for solid tumors (Vogelstein et al., 2013).

## 2.4 Results

We calculated the proportion of drive gene mutations caused by E or H in 32 cancer types. We considered those mutations not attributable to either E or H to be due to R. Figure 2.1 and Figure 2.2 show our estimates for females and males respectively. When looking across the 32 cancer types, the median proportion of driver mutations attributable to E was 23%. The estimate varied considerably: it was greater than 60% in cancers such as those of the lung,

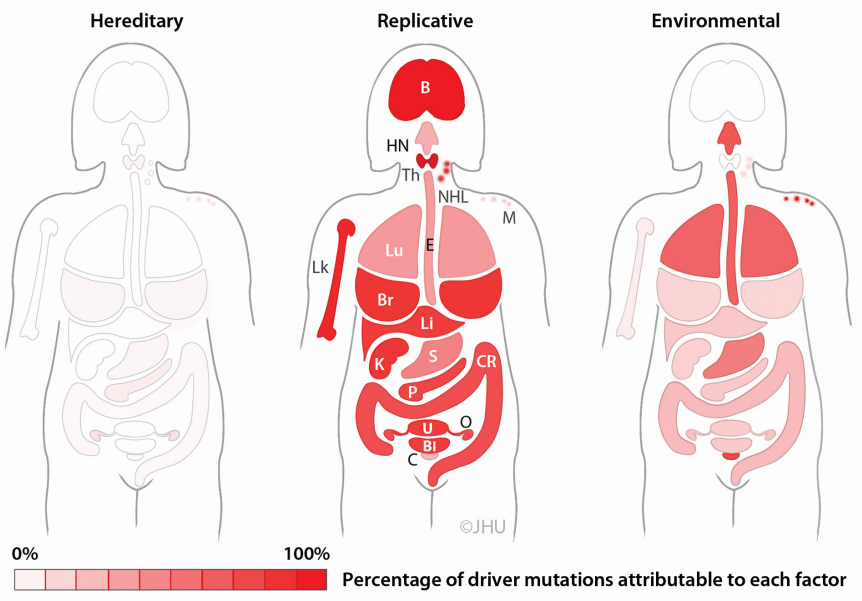


**Figure 2.1: Etiology of driver gene mutations in women with cancer** For each of 18 representative cancer types, the schematic depicts the proportion of mutations that are inherited, due to environmental factors, or due to errors in DNA replication (i.e., not attributable to either heredity or environment). The sum of these three proportions is 100%. The color codes for hereditary, replicative, and environmental factors are identical and span white (0%) to brightest red (100%). B, brain; Bl, bladder; Br, breast; C, cervical; CR, colorectal; E, esophagus; HN, head and neck; K, kidney; Li, liver; Lk, leukemia; Lu, lung; M, melanoma; NHL, non-Hodgkin lymphoma; O, ovarian; P, pancreas; S, stomach; Th, thyroid; U, uterus.

esophagus, and skin and 15% or less in cancers such as those of the prostate, brain and breast. When normalized for the incidence of each of these 32 cancer types, we calculated that 29% of the mutations in cancers occurring in the UK were attributable to E, 5% of the mutations were attributable to H, and 66% were presumably attributable to R. The numerical results and the values used to construct Figures 2.1 and 2.2 are provided in Table 2.1. In addition, UK Cancer Research estimates that 42% of these cancer cases are preventable. Given the relationship between PAR and $P_E$ we provided above, the proportion of mutations caused by environmental factors is always less

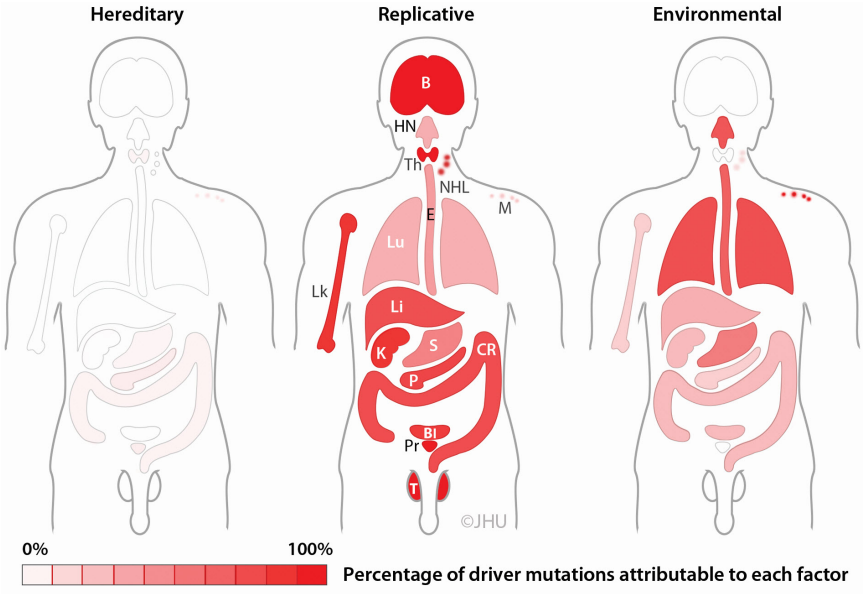


**Figure 2.2: Proportion of driver gene mutations attributable to H, R, and E in men with cancer** For each of 16 representative cancer types, the schematic depicts the proportion of mutations that are inherited, due to environmental factors, or due to errors in DNA replication (i.e., not attributable to either heredity or environment). The sum of these three proportions is 100%. The color codes for hereditary, replicative, and environmental factors are identical and span white (0%) to brightest red (100%). B, brain; Bl, bladder; CR, colorectal; E, esophagus; HN, head and neck; K, kidney; Li, liver; Lk, leukemia; Lu, lung; M, melanoma; NHL, non-Hodgkin lymphoma; P, pancreas; Pr, prostate; S, stomach; T, testis; Th, thyroid; U, uterus.

than the proportion of cancers preventable by avoiding these factors. Thus, our estimate that a maximum of 29% of the mutations in these cancers are due to E is perfectly compatible with the estimate that 42% of these cancers are preventable, in fact they were based on those epidemiological estimates.

## 2.5 Discussion

The results described above have important ramifications for understanding the root causes of cancer as well as for minimizing deaths from this disease.

Our approach - a combination of cancer sequencing data and conservative analyses of environmental and hereditary risk factors - provides the estimates of the actual contribution of R mutations to a certain cancer type. They indicated that even in lung adenocarcinomas, R contributes a third of the total mutations, with tobacco smoke (including second-hand smoking), diet, radiation, and occupational exposures contributing the remainder. In cancers that are less strongly associated with environmental factors, such as those of the pancreas, brain, bone, or prostate, the majority of the mutations are attributable to R.

Our results explicitly and quantitatively address the difference between cancer etiology and cancer preventability. As illustrated in Figures 2.1 and 2.2, these concepts are not equivalent. A cancer in which 50% of the mutations are due to R can still be preventable. The reason for this is that it generally requires more than one mutation to develop the disease. A cancer that required two mutations is still preventable if one of the mutations was due to R and the other due to an avoidable environmental factor.

Our results are fully consistent with epidemiological evidence on the fraction of cancers in developed countries that are potentially preventable through improvements in environment and lifestyle. Cancer Research UK estimates that 42% of cancer cases are preventable (Cancer Research UK, 2018b); the U.S. Centers for Disease Control and Prevention estimates that 21% of annual cancer deaths in individuals less than 80 years old could be prevented (US Centers for Disease Control and Prevention, 2014).

Of equal importance, these studies provide a well-defined, molecular

explanation for the large and apparently unpreventable component of cancer risk that has long puzzled epidemiologists. It is, of course, possible that virtually all mutations in all cancers are due to environmental factors, most of which have simply not yet been discovered. However, such a possibility seems inconsistent with the exhaustively documented fact that about three mutations occur every time a normal cell divides and that normal stem cells often divide throughout life.

Our studies complement, rather than oppose, those of classic epidemiology. For example, the recognition of a third, major factor (R) underlying cancer risk can inform epidemiologic studies by pointing to cancers that cannot yet be explained by R (i.e., those with too few stem cell divisions to account for cancer incidence). Such cancer types seem particularly well suited for further epidemiologic investigation. Additionally, R mutations appear unavoidable now, but it is conceivable that they will become avoidable in the future. There are at least four sources of R mutations in normal cells: quantum effects on base pairing (Kimsey et al., 2015), mistakes made by polymerases (Kunkel, 2009), hydrolytic deamination of bases (Fromme and Verdine, 2004), and damage by endogenously produced reactive oxygen species or other metabolites (Collins, 2005). The last of these could theoretically be reduced by the administration of antioxidant drugs (Ferguson et al., 2015). The effects of all four could, in principle, be reduced by introducing more efficient repair genes into the nuclei of somatic cells or through other creative means.

As a result of the aging of the human population, cancer is today the most common cause of death in the world (Stewart and Wild, 2014). Primary

prevention is the best way to reduce cancer deaths. Recognition of a third contributor to cancer - R mutations - does not diminish the importance of primary prevention but emphasizes that not all cancers can be prevented by avoiding environmental risk factors (Figures 2.1 and 2.2). Fortunately, primary prevention is not the only type of prevention that exists or can be improved in the future. Secondary prevention, i.e., early detection and intervention, can also be lifesaving. For cancers in which all mutations are the result of R, secondary prevention is the only option.

| cancer type | female_E | female_H | female_R | male_E | male_H | male_R | both_E | both_H | both_R |
|---|---|---|---|---|---|---|---|---|---|
| anal | 0.6 | 0.005 | 0.395 | 0.6 | 0.005 | 0.395 | 0.6 | 0.005 | 0.395 |
| bladder | 0.191 | 0.005 | 0.804 | 0.252 | 0.005 | 0.743 | 0.235 | 0.005 | 0.76 |
| bone | 0 | 0.005 | 0.995 | 0 | 0.005 | 0.995 | 0 | 0.005 | 0.995 |
| brain | 0.001 | 0.005 | 0.994 | 0003 | 0.005 | 0.992 | 2E-3 | 0.005 | 0.993 |
| breast | 0.151 | 0.015 | 0.834 | NA | NA | NA | 0.151 | 0.015 | 0.834 |
| cervix | 0.746 | 0.005 | 0.249 | NA | NA | NA | 0.746 | 0.005 | 0.249 |
| colon | 0.249 | 0.025 | 0.726 | 0.27 | 0.025 | 0.705 | 0.261 | 0.025 | 0.714 |
| esophagus | 0.628 | 0.005 | 0.367 | 0.596 | 0.005 | 0.399 | 0.606 | 0.005 | 0.389 |
| eye | 0.003 | 0.2 | 0.797 | 0.021 | 0.2 | 0.779 | 0.012 | 0.2 | 0.788 |
| gallbladder | 0.079 | 0.005 | 0.916 | 0.08 | 0.005 | 0.915 | 0.079 | 0.005 | 0.916 |
| Hodgkin | 0.3 | 0.005 | 0.695 | 0.3 | 0.005 | 0.695 | 0.3 | 0.005 | 0.695 |
| kidney | 0.159 | 0.01 | 0.831 | 0.236 | 0.01 | 0.754 | 0.207 | 0.01 | 0.783 |
| larynx | 0.693 | 0.005 | 0.302 | 0.716 | 0.005 | 0.279 | 0.709 | 0.005 | 0.286 |
| leukemia | 0.106 | 0.005 | 0.889 | 0.168 | 0.005 | 0.827 | 0.143 | 0.005 | 0.852 |
| liver | 0.199 | 0.005 | 0.796 | 0.287 | 0.005 | 0.708 | 0.257 | 0.005 | 0.738 |
| lung | 0.613 | 0.005 | 0.382 | 0.703 | 0.005 | 0.292 | 0.661 | 0.005 | 0.334 |
| melanoma | 0.86 | 0.05 | 0.09 | 0.86 | 0.05 | 0.09 | 0.86 | 0.05 | 0.09 |
| mesothelioma | 0.525 | 0.005 | 0.47 | 0.685 | 0.005 | 0.31 | 0.657 | 0.005 | 0.338 |
| myeloma | 0.001 | 0.005 | 0.994 | 0.003 | 0.005 | 0.992 | 0.003 | 0.005 | 0.993 |
| non-Hodgkin | 0.035 | 0.005 | 0.96 | 0.043 | 0.005 | 0.952 | 0.039 | 0.005 | 0.956 |
| oral | 0.68 | 0.005 | 0.315 | 0.721 | 0.005 | 0.274 | 0.708 | 0.005 | 0.287 |
| ovary | 0.221 | 0.05 | 0.729 | NA | NA | NA | 0.221 | 0.05 | 0.729 |
| pancreas | 0.187 | 0.05 | 0.763 | 0.172 | 0.05 | 0.778 | 0.18 | 0.05 | 0.77 |
| penile | NA | NA | NA | 0.267 | 0.005 | 0.728 | 0.267 | 0.005 | 0.728 |
| prostate | NA | NA | NA | 0 | 0.045 | 0.955 | 0 | 0.045 | 0.955 |
| stomach | 0.509 | 0.015 | 0.476 | 0.576 | 0.015 | 0.409 | 0.553 | 0.015 | 0.432 |
| testicular | NA | NA | NA | 0 | 0.005 | 0.995 | 0 | 0.005 | 0.995 |
| thyroid | 0.005 | 0.015 | 0.98 | 0.004 | 0.015 | 0.981 | 0.005 | 0.015 | 0.98 |
| uterus | 0.181 | 0.01 | 0.809 | NA | NA | NA | 0.181 | 0.01 | 0.809 |
| vaginal | 0.42 | 0.005 | 0.575 | NA | NA | NA | 0.42 | 0.005 | 0.575 |
| vulval | 0.267 | 0.005 | 0.728 | NA | NA | NA | 0.267 | 0.005 | 0.728 |
| others | 0.057 | 0.005 | 0.938 | 0.064 | 0.005 | 0.931 | 0.057 | 0.005 | 0.938 |
| oveall | 0.278 | 0.05 | 0.672 | 0.299 | 0.05 | 0.651 | 0.289 | 0.05 | 0.661 |

**Table 2.1: Proportion of driver gene mutations attributable to E, H, and R** Col 1: Cancer type; Col 2-4: Female; Col 5-7: Male; Col 8-10: Both sexes. The values listed in columns 2, 5, and 8 represent the estimated fraction of the driver gene mutations attributable to environmental factors, $P_E$, for each indicated cancer type. The values listed in columns 3, 6, and 9 are the estimated fraction of the driver gene mutations attributable to inherited factors, $P_H$, for each indicated cancer type. The values listed in columns 4, 7, and 10 are the estimated fraction of the driver gene mutations attributable to replicative mutations, $P_R$, for each indicated cancer type. The overall values represents the risks for all cancer types after normalizing for cancer incidence, so that the higher the incidence of the cancer, the more weight it is given.

# References


Garraway, Levi A and Eric S Lander (2013). "Lessons from the cancer genome". In: *Cell* 153.1, pp. 17–37.

Stratton, Michael R, Peter J Campbell, and P Andrew Futreal (2009). "The cancer genome". In: *Nature* 458.7239, p. 719.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler (2013). "Cancer genome landscapes". In: *science* 339.6127, pp. 1546–1558.

Mucci, Lorelei A, Jacob B Hjelmborg, Jennifer R Harris, Kamila Czene, David J Havelick, Thomas Scheike, Rebecca E Graff, Klaus Holst, Sören Möller, Robert H Unger, et al. (2016). "Familial risk and heritability of cancer among twins in Nordic countries". In: *Jama* 315.1, pp. 68–76.

Stadler, Zsofia K, Peter Thom, Mark E Robson, Jeffrey N Weitzel, Noah D Kauff, Karen E Hurley, Vincent Devlin, Bert Gold, Robert J Klein, and Kenneth Offit (2010). "Genome-wide association studies of cancer". In: *Journal of Clinical Oncology* 28.27, p. 4255.

Tomasetti, Cristian and Bert Vogelstein (2015b). "Variation in cancer risk among tissues can be explained by the number of stem cell divisions". In: *Science* 347.6217, pp. 78–81.

Tomasetti, Cristian and Bert Vogelstein (2015a). "Musings on the theory that variation in cancer risk among tissues can be explained by the number of divisions of normal stem cells". In: *arXiv preprint arXiv:1501.05035*.

Lynch, Michael (2010). "Rate, molecular spectrum, and consequences of human mutation". In: *Proceedings of the National Academy of Sciences* 107.3, pp. 961–968.

Tomasetti, Cristian, Bert Vogelstein, and Giovanni Parmigiani (2013). "Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation". In: *Proceedings of the National Academy of Sciences* 110.6, pp. 1999–2004.

Zhu, Liqin, David Finkelstein, Culian Gao, Lei Shi, Yongdong Wang, Dolores López-Terrada, Kasper Wang, Sarah Utley, Stanley Pounds, Geoffrey Neale, et al. (2016). "Multi-organ mapping of cancer risk". In: *Cell* 166.5, pp. 1132–1146.

Cancer Research UK (2018b). *Preventable cancers*. URL: www.cancerresearchuk.org/health-professional/cancer-statistics/risk/preventable-cancers (visited on 10/20/2018).

Tomasetti, Cristian, Luigi Marchionni, Martin A Nowak, Giovanni Parmigiani, and Bert Vogelstein (2015). "Only three driver gene mutations are required for the development of lung and colorectal cancers". In: *Proceedings of the National Academy of Sciences* 112.1, pp. 118–123.

Durrett, Richard and Stephen Moseley (2010). "Evolution of resistance and progression to disease during clonal expansion of cancer". In: *Theoretical population biology* 77.1, pp. 42–48.

McLaughlin-Drubin, Margaret E, Jordan Meyers, and Karl Munger (2012). "Cancer associated human papillomaviruses". In: *Current opinion in virology* 2.4, pp. 459–466.

Cancer Research UK (2018c). *Statistics by cancer type*. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type (visited on 10/20/2018).

Parkin, DM (2011). "1. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010". In: *British journal of cancer* 105.S2, S2.

UK Office for National Statistics (2015). *Population estimates for UK, England and Wales, Scotland and Northern Ireland: Mid-2015*. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2015 (visited on 10/20/2018).

Cancer Research UK (2018a). *Inherited cancer genes and increased cancer risk*. URL: https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/inherited-cancer-genes-and-increased-cancer-risk/inherited-genes-and-cancer-types-inherited_genes4 (visited on 10/20/2018).

National Cancer Institute (2018). *Genetic testing for hereditary cancer syndromes*. URL: https://www.cancer.gov/about-cancer/causes-prevention/genetics/genetic-testing-fact-sheet (visited on 10/20/2018).

US Centers for Disease Control and Prevention (2014). *Up to 40 percent of annual deaths from each of five leading US causes are preventable*. URL: https:

//www.cdc.gov/media/releases/2014/p0501-preventable-deaths.html (visited on 10/20/2018).

Kimsey, Isaac J, Katja Petzold, Bharathwaj Sathyamoorthy, Zachary W Stein, and Hashim M Al-Hashimi (2015). "Visualizing transient Watson–Crick-like mispairs in DNA and RNA duplexes". In: *Nature* 519.7543, p. 315.

Kunkel, Thomas A (2009). "Evolving views of DNA replication (in) fidelity". In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 74. Cold Spring Harbor Laboratory Press, pp. 91–101.

Fromme, J Christopher and Gregory L Verdine (2004). "Base excision repair". In: *Advances in protein chemistry*. Vol. 69. Elsevier, pp. 1–41.

Collins, Andrew R (2005). "Antioxidant intervention as a route to cancer prevention". In: *European Journal of Cancer* 41.13, pp. 1923–1930.

Ferguson, Lynnette R, Helen Chen, Andrew R Collins, Marisa Connell, Giovanna Damia, Santanu Dasgupta, Meenakshi Malhotra, Alan K Meeker, Amedeo Amedei, Amr Amin, et al. (2015). "Genomic instability in human cancer: Molecular insights and opportunities for therapeutic attack and prevention through diet and nutrition". In: *Seminars in cancer biology*. Vol. 35. Elsevier, S5–S24.

Stewart, Bernard W and CP Wild (2014). "World Cancer Report 2014. Lyon, France: International Agency for Research on Cancer". In: *World Health Organization*, p. 630.

# Chapter 3

# Impact of somatic mutation rate in cancer etiology

## 3.1 Introduction

Understanding the etiology of cancer is critical for both its prevention and its cure (Song et al., 2018). Large scientific evidence has pointed for decades to the role played by environmental (E) and inherited (H) factors (Cancer Research UK, 2018; Health, Services, et al., 2004). Recent research has added a previously unappreciated third factor: the replicative mutations (R) that accumulate normally in our tissues every time a cell divides (Tomasetti and Vogelstein, 2015; Tomasetti, Li, and Vogelstein, 2017). It has been estimated that the majority of the mutations found in cancers are due to this R factor, when weighting cancer types by their incidence (Tomasetti, Li, and Vogelstein, 2017).

However, independently of whether these mutations are contributed by E, H, or R, a fundamental but completely unanswered question remains: how much of cancer is due to mutations?

The role of driver mutations and other genomic alterations in cancer causation is well recognized, and much experimental evidence points to their step-wise accumulation. Thus, mutations certainly represent a necessary ingredient for cancer to occur. At the same time, they may not be sufficient, that is, it is not known how large is their contribution when compared to other important factors like, for example, the immune system and the microenvironment. In fact the case may be made that, with aging, several cells will have acquired the necessary mutations, but only a few or possibly none of them will end up yielding a cancer. Only those able to overcome the strictly regulated conditions dictated by their microenvironment, as well as able to escape the immune system and its protective functions, will. And both E and H may play a direct role in limiting the ability of the immune system and the tissue's microenvironment to provide their functions.

It is then natural to ask how large is the role of driver mutations in cancer causation when compared to factors like the immune systems, the microenvironment, and many other, possibly unknown, ones. More precisely, can we assess quantitatively how much of the effects of each of the three etiological factors, E, H, and R, occur through mutations versus other mechanisms, which we collectively define as K factors? It is clear that the R acts exclusively through mutations, by its very definition, but what about E and H factors? For example, is smoking - a well-known E factor - increasing cancer risk by simply inducing a higher mutation rate, or rather by damaging the immune system and other K factors, or both? And if both, is it possible to estimate those proportions?

This study represents the first attempt to assess this question. The overall idea is that we can leverage our knowledge of the effect that a given increase in the background mutation rate will have in increasing cancer risk. Since we can estimate the increase in the mutation rate caused by an H or E factor, via the analysis of sequencing data, we can then estimate the expected increased risk for a given cancer type, if that H or E factor acted only through its effects on mutations. The comparison of that expected increase in risk with the epidemiologically observed relative risk (RR) induced by that E or H factor will provide then an estimate of the proportion of the actual relative risk attributable simply to the effects of mutations.

## 3.2 Method

The probability for a person to develop cancer can be thought of as a function of multiple factors, including but not limited to, the number of stem cell divisions, age, the mutation probability per cell division, the immune system and the microenvironment. Here, we define the mutation rate $u$ in a stem cell lineage of a given tissue as the number of somatic mutations occurred in one year in that stem cell lineage. To investigate the role of the mutation rate in cancer etiology, we propose the following method.

Based on previous results (Tomasetti, Vogelstein, and Parmigiani, 2013; Tomasetti et al., 2015), the mutation rate $u$ has a multiplicative effect on $P$, which is defined as the probability of getting a given cancer for an individual by age $t$. Considering $P$ as a function of age, the mutation rate, and all other

relevant factors, we have

$$P(t,u) = u^n f(\#\text{stem cells}, \text{t}, \text{immune system}, \text{microenvironment}, \text{etc.}) \quad (3.1)$$

where $t$ and $u$ are the age and the mutation rate of that individual, $n$ is the power-law effect of the mutation rate on cancer risk depending on the number of drivers required to yield a certain cancer, and $f$ is just the symbol for some unknown function of all the arguments inside the parenthesis. We note that $n$ does not have to be an integer. To provide some examples of the possible values for $n$, and by assuming that each successive driver gene mutation doubled the proliferative fitness advantage (Tomasetti et al., 2015; Durrett and Moseley, 2010), if 2 drivers are required to get to cancer, then $n = 2$; if 3 drivers are required, then $n = 2.5$. While we do not know the exact value of $n$, it is easy to show that for solid tumors $n$ should range between 2 and 3 (Tomasetti et al., 2015). In fact, it has been shown that 2.5 is a good approximation and certainly the case $n = 2$ is conservative with respect to our goal of estimating the contribution of the increase in mutations in explaining relative risk (see below).

Let $K$ represent the overall average effect on cancer risk other than the mutation rate, and considering $t$ as the lifespan of an individual, Equation 3.1 can be written as

$$P(u) = u^n K \quad (3.2)$$

Equation 3.2 represents the lifetime risk of getting a certain cancer for an individual among the population unexposed to any extrinsic factors (unexposed population) with an individual-specific mutation rate $u$. For people who are

exposed to an extrinsic risk factor $E$ (exposed population), like smoking, their lifetime risk is

$$P(u) = u^n K^* \tag{3.3}$$

where $K^*$ is accounting for all the non-mutational factors which affect risk: the original ones that were already included in $K$, plus any additional effects added by the exposure to the $E$ factor. By writing it in this form, we are able to separate cancer risk into mutational and non-mutational effects. It can be inferred that if $K^* \approx K$, then the risk factor $E$ is mostly acting via increasing the mutation rate, that is by shifting to the right in which case the increase in mutation rate should be able to explain most of the increased cancer risk.

Assume that $g(u)$ is the density function of the distribution of the mutation rate among the unexposed people in the general population (not restricted to cancer patients), then

$$LRC = \int_0^{+\infty} P(u)g(u)du \tag{3.4}$$

where $LRC$ is the lifetime risk of getting a certain type of cancer for the unexposed people. Similarly, $LRC^*$ for those in the general population exposed to the $E$ or $H$ factor, with density $g^*(u)$ for the mutation rate can be written as

$$LRC^* = \int_0^{+\infty} P^*(u)g^*(u)du \tag{3.5}$$

Let $g_c(u)$ be the density function of the distribution of the mutation rate among the unexposed cancer patients. This can be thought of as a conditional density of $u$, i.e. conditional on having cancer, and its relationship with $g(u)$, the unconditional density, is given by $g_c(u) = \frac{P(u)g(u)}{LRC}$. Therefore, based on

Equation 3.2, we have

$$g(u) = g_c(u)\frac{LRC}{u^n K} \tag{3.6}$$

Since $g(u)$ is a density function, $\int_0^{+\infty} g(u)du = 1$. Taking the integral from both sides in Equation 3.6 with respect to $u$

$$1 = \int_0^{+\infty} g_c(u)\frac{LRC}{u^n K}du \tag{3.7}$$

Thus,

$$K = LRC\int_0^{+\infty} g_c(u)u^{-n}du \tag{3.8}$$

and similarly,

$$K^* = LRC^*\int_0^{+\infty} g_c^*(u)u^{-n}du \tag{3.9}$$

One major advantage of using the conditional density to estimate $K$ and $K^*$ is that we can use the mutation data from large sequencing studies on cancer patients, while the information on the whole healthy population is in general not available. As mentioned earlier, $K$ and $K^*$ are the average non-mutational effects on cancer risks. The relative relationship between $K^*$ and $K$ serves as a good estimate of the increased risk due to the effects of the $E$ or $H$ factor via mechanisms other than the mutation rate. Now, the relative risk (RR) of a risk factor $E$ can be expressed using the ratio of LRC's. Using Equation 3.8 and 3.9, we have

$$RR = \frac{LRC^*}{LRC} = \frac{K^*}{K}\frac{\int_0^{+\infty} g_c(u)u^{-n}du}{\int_0^{+\infty} g_c^*(u)u^{-n}du} \triangleq \mathcal{KU} \tag{3.10}$$

The relative risk induced by the $E$ or $H$ factor is therefore split into the two possible pathways through which it can act: mutational ($\mathcal{U}$) and non-mutational effects ($\mathcal{K}$). $\mathcal{U}$ and $\mathcal{K}$ have multiplicative effects on the increased cancer risk.

We calculate $p_u$, the proportion of increased risk due to the effect on the mutation rate, using the following formula:

$$p_u = \frac{\mathcal{U} - 1}{RR - 1} \tag{3.11}$$

The numerator is the excess relative risk purely due to the increase in the mutation rate induced by the $E$ or $H$ factor, while the denominator is the total excess relative risk. On the other hand, we could also calculate $p_u$ by subtracting the proportion due to the increase of non-mutational effects caused by $E$ or $H$ factor:

$$p_u = 1 - \frac{\mathcal{K} - 1}{RR - 1} \tag{3.12}$$

The definition in Equation 3.11 ignores the interaction effects of $\mathcal{U}$ and $\mathcal{K}$. While in Equation 3.12, all the interaction effects are attributed to mutations. Thus, we define $p_u$ as the average from these two expressions.

To estimate $p_u$ from the data, the following method based on the Appoximate Bayesian Computation method (Beaumont, Zhang, and Balding, 2002) was applied in order to obtain confidence intervals for $K^*/K$, as well as for the proportion of increased risk due to the increase in the mutation rate induced by the $E$ or $H$ factor. The method is described as follows.

Step 1: Sample $M$ values of $K^*/K$ from a prior distribution, in our case a uniform distribution that depends on the $RR$ for each specific risk factor.

Step 2: For each $K^*/K$, using bootstrap to get replicates of the distribution of $u$ from the sequencing data set, calculate $\mathcal{U} \triangleq \frac{\int_0^{+\infty} g_c(u)u^{-n}du}{\int_0^{+\infty} g_c^*(u)u^{-n}du}$ for each replicate and obtain the corresponding $RR$ using Equation 3.10.

Step 3: If the median $RR$ of all replicates with the same $K^*/K$ is within the 95% confidence interval of the $RR$ from epidemiological studies, then accept this value of $K^*/K$; otherwise, reject.

Step 4: All accepted values form the posterior distribution of $K^*/K$, which is used to obtain the empirical 95% confidence interval for $K^*/K$ and $p_u$.

## 3.3 Data

Cancer genome sequencing data from The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas, 2018) was used for this analysis. We investigated 4 known risk factors associated with 7 cancer types. There is information on several other factors/cancer types in the TCGA database. However, since our method requires density estimations, we excluded from the analysis factors/cancer types that did not have enough patients out of accuracy concerns.

Risk factors included in this study were smoking, BMI, microsatellite instability (MSI), and Hepatitis C (HCV) infection. Cancer types included lung, kidney, head and neck, bladder, liver, uterus, and colon cancers. Patients with unknown age information were excluded. Before analyzing the data, some filters were applied to account for potential sequencing errors. Standard filters were at least 2 but no more than 1000 mutations in total, not higher than 10 times the median mutation rate in the corresponding group. Patients with BMI $\leq$ 27.5 were defined as normal in order to increase the sample size in the normal group.

## 3.4 Results

### 3.4.1 Smoking

#### 3.4.1.1 Lung Adenocarcinoma (LUAD)

TCGA-LUAD data was used to perform the analysis. After removing all ineligible cases, 62 nonsmokers and 383 smokers were qualified for our analysis. Comparing to nonsmokers, current male and female smokers are about 10.8 (95% CI 8.7-13.3) and 4.2 (95% CI 3.5-5) times more likely to develop lung adenocarcinoma respectively (Pesch et al., 2012). This gives an approximate 7.45 (95% CI 6.06-9.08) times higher risk for current smokers based on the gender distribution in our data. $K^*/K$ is estimated to be 1.50 or 1.39 with $n = 2$ or 2.5, yielding the proportions of relative risk in lung cancer explained by the effect of smoking on the mutation rate as 76.78% (95% CI 68.17%-86.80%) with $n = 2$ and 80.71% (95% CI 69.85%-92.01%) with $n = 2.5$.

#### 3.4.1.2 Kidney Renal Cell Carcinoma (RCC)

Since smoking is a risk factor for clear cell and papillary renal cell carcinoma (Patel et al., 2015), we used these two subtypes data (TCGA-KIRC and KIRP) in the analysis. 61 nonsmokers and 51 smokers were analyzed correspondingly. It is estimated that the increased risk for kidney cancer due to smoking is 54% (95% CI 42%-68%)) in males and 22% (95% CI: 9%-36%) in females (Hunt et al., 2005). The relative risk, averaged by genders in the data, is 1.44 (95% CI: 1.32-1.58). The ratio between $K^*$ and $K$, $K^*/K$, is estimated to be 0.99 or 0.91 with $n = 2$ or 2.5, indicating that the increased mutation rate is able to explain all the increased cancer risk. $p_u$ is 100% in both cases,

with confidence intervals (70.90%, 100%) and (86.36%,100%) for $n = 2, 2.5$ respectively.

#### 3.4.1.3 Head and Neck Squamous Cell Carcinoma (HNSC)

The majority of head and neck cancers are squamous cell carcinomas. In terms of the higher risk due to smoking, previous studies have showed that current smokers have 1.91 (95% CI 1.06-3.42) times higher risk in oral cavity cancer, 7.49 (95% CI 2.87-19.54) times higher risk in pharyngeal cancer, and 5.26 (95% CI 2.45-11.28) times higher risk in laryngeal cancer (Maasland et al., 2014). To control for the effects of HPV, this analysis was restricted on HPV-negative cases. 47 nonsmokers and 216 smokers were finally included. $K*/K$ is estimated to be 1.2190 and 0.9087 with $n = 2$ and 2.5. $p_u$ is therefore 82.95% (95% CI 50.48%-100%) and 100% (95% CI 62.66%-100%) for $n = 2$ and 2.5 respectively.

#### 3.4.1.4 Bladder Urothelial Carcinoma (BLAC)

Urothelial bladder carcinoma is the most common type of bladder cancer. Previous studies suggested a relative risk of 4.06 (95%CI 3.66-4.50) to develop bladder cancer when comparing smokers with nonsmokers (Freedman et al., 2011). 108 nonsmokers and 279 smokers were included in the analysis. $K^*/K$ is estimated to be 1.93 (n=2) and 1.48 (n=3). $p_u$ is then 52.81% (95% CI 45.54%-58.01%) assuming $n = 2$ and 70.72% (95% CI 60.51%-78.62%) assuming $n = 2.5$.

### 3.4.2 Body Mass Index (BMI)

#### 3.4.2.1 Uterine Corpus Endometrial Carcinoma (UCEC)

The analysis was restricted on patients with uterus cancers who are Microsatellite Stable (MSS). As a result, 74 normal and 204 obese people were included in the analysis. Previous study suggested that the relative risk for uterus cancer is 1.5 (95%CI 1.26-1.78) for BMI from 25-29.9, 2.53 (95%CI 2.02-3.18) for BMI 30.0-34.9, 2.77 (95%CI 1.83-4.18) for BMI 35.0-39.9, and 6.25 (95%CI 3.75-10.42) for BMI of 40 and above (Calle et al., 2003). This suggests a relative risk of 3.04 (95% CI 2.20-4.24) based on the BMI distribution in our data. The calculated $K^*/K$ is 2.62 (n=2) or 2.34 (n=2.5). As a result, $p_u$ is 14.15% (95% CI 8.48%-18.58%) assuming $n = 2$ or 24.51% (95% CI 15.55%-32.44%).

#### 3.4.2.2 Colon Adenocarcinoma (COAD)

The analysis was performed on microsatellite stable colorectal cancer cases. Previous study suggests that for female, the relative risk for colon cancer is 1.10 (95%CI 1.01-1.19) for BMI 25.0-29.9, 1.33 (95%CI 1.17-1.51) for BMI 30.0-34.9, 1.36 (95%CI 1.06-1.74) for BMI 35.0-39.9, and 1.46 (95%CI 0.94-2.24) for BMI of 40 and above; for male, it is 1.2 (95%CI 1.12-1.30) for BMI 25.0-29.9, 1.47 (95%CI 1.30-1.66) for BMI 30.0-34.9, and 1.84 (95%CI 1.39-2.41) for BMI 35 and above ((Calle et al., 2003)). There were 71 normal and 74 obese cases, and the ratio $K^*/K$ is estimated to be 1.41 (n=2) or 1.44 (n=2.5). What is the BMI distribution in our data? This indicates that the mutation rate explain essentially none on the increased cancer risk for colorectal cancer due to obesity.

### 3.4.3 Virus: Hepatitis C (HCV)

#### 3.4.3.1 Liver Hepatocellular Carcinoma (LIHC)

Since the patients in the HCV positive group are all alcohol positive, we used only alcohol positive within the HCV negative group to control for the factor of drinking alcohol. Therefore, 74 HCV negative and 153 HCV positive cases were included in the analysis. Previous study suggested that the relative risk for liver cancer due to HCV is 33.3 (95% 13.6-81.3) for female, and 17.6 (95% CI 10.7-28.8) for male (Donato et al., 2002). The overall relative risk based on the gender distribution in the data is 18.75. $K^*/K$ is estimated to be 5.83 (n=2) or 3.73 (n=2.5). As a result, $p_u$ is 42.65% (95% CI 38.45%-48.20%) with $n = 2$ or 53.68% (95% CI 47.01%-62.84%) with $n = 2.5$.

### 3.4.4 Microsatellite instability (MSI)

#### 3.4.4.1 Uterine Corpus Endometrial Carcinoma (UCEC)

Women with mismatch repair deficiency are reported to have a 76.0 (95% CI 56.3-102.5) times the risk of developing endometrial cancer (Stoffel et al., 2009). To control for the effect of BMI, we only included patients with BMI<=27.5 for this analysis. Microsatellite instability-high (MSI-h) cancers were defined as the MSI group. 74 MSS and 39 MSI cases were used to perform the analysis, which estimates $K^*/K$ as 1.37 (n=2) or 0.65 (n=2.5). Correspondingly, $p_u$ is 86.11% (95% CI 76.87%-100%) with $n = 2$ or 100% with $n = 2.5$. This means almost all increased risk caused by MSI is due to an increased mutation rate.

#### 3.4.4.2 Colon Adenocarcinoma (COAD)

People with mismatch repair deficiency have significant higher risk of developing colon cancer. The hazard ratio for males is 262.7 (95% CI 214.7-321.5), for females is 128.0 (95% CI 97.6-168.0) (Stoffel et al., 2009). Only microsatellite instability-high (MSI-h) cases were included in MSI group. As a result, 232 MSS and 66 MSI observations were qualified for the analysis. Considering the gender distribution in the data, the estimated ratio of $K^*$ and K is 1.38 (n=2) and 0.38 (n=2.5). $p_u$ is then 86.05% (95% CI 78.94%-97.12%) assuming $n = 2$ or 100% assuming $n = 2.5$.

## 3.5 Discussion

Our study is the first one to quantitatively assess the proportion of the relative risk caused by an $E$ or $H$ factor that can be explained by their increase of the mutation rate. The results described above have important indications on the role of driver mutations in cancer causation.

Among the 4 factors that we have analyzed, smoking and MSI are the two where an increased mutation rate is able to explain the majority, if not all, of the increased cancer risk. The results with respect to smoking are not surprising since scientists have known for decades that smoking cigarettes causes DNA damage (Pfeifer et al., 2002). And the result for MSI are precisely what expected, as patients with mismatch repair deficiency are much more likely to get cancer because they accumulate significantly more mutations than normal people.

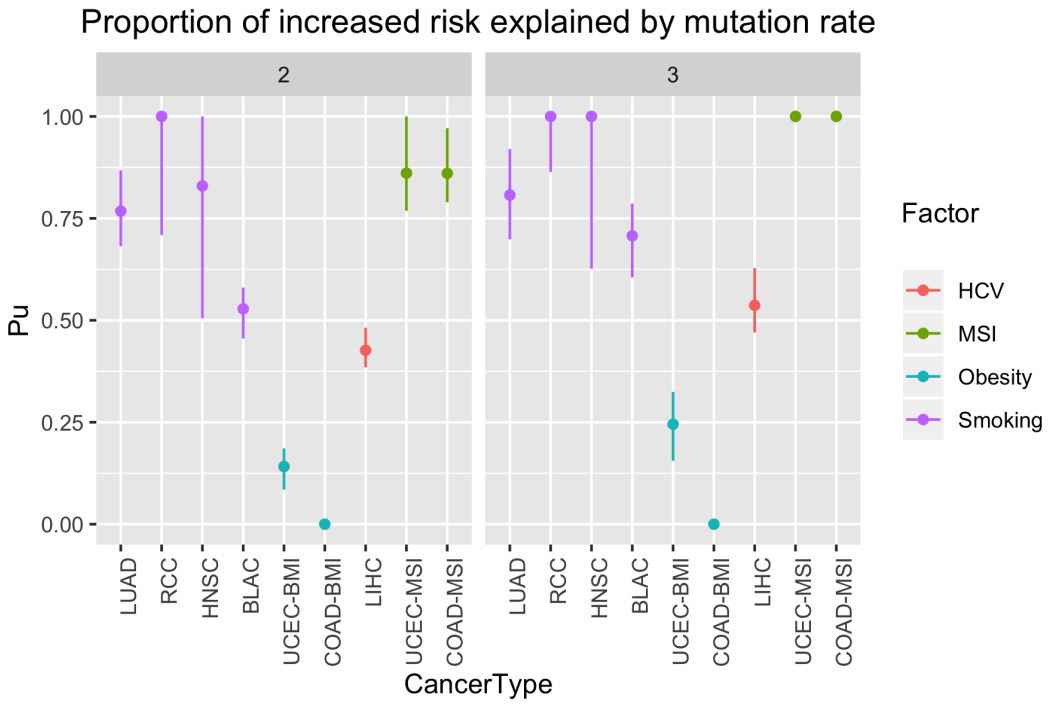


**Figure 3.1: bf Proportion of relative risk explained by the increase of the mutation rate caused by the indicated $E$ or $H$ factor.** Estimate of $P_u$ by assuming that 2 or 3 driver mutations are required for developing cancer, respectively. In smoking and MSI, the observed increase in the mutation rate is able to explain the majority if not all of the increased cancer risk. While for BMI, the mutation rate can explain little about the higher cancer relative risk due to obesity. Cancer types: LUAD (Lung Adenocarcinoma); RCC (Kidney Renal Cell Carcinoma); HNSC (Head and Neck Squamous Cell Carcinoma); BLAC (Bladder Urothelial Carcinoma); UCEC-BMI (Uterine Corpus Endometrial Carcinoma considering BMI); COAD-BMI (Colon Adenocarcinoma considering BMI); LIHC (Liver Hepatocellular Carcinoma); UCEC-MSI (Uterine Corpus Endometrial Carcinoma considering MSI); COAD-MSI (Colon Adenocarcinoma considering MSI)

The increase in background mutation rate can also partially explain the higher cancer risk in HCV-positive population. This is also expected when we look at the mechanisms of carcinogenesis in HCV-associated liver cancers. Chronic HCV infections are typically associated with inflammatory responses within the liver, leading to progressive fibrosis and cirrhosis. This, and other evidence suggesting that HCV can disrupt the control of cellular proliferation, provide an important mechanism for carcinogenesis, since increasing the rate of cell division increases the mutation rate. Moreover, evidence shows that HCV can impair the cell's response to DNA damage (McGivern and Lemon, 2011). This mutational effect may "hide" some of the mutational consequences of an HCV infection in causing liver cancer, since the automatic acquisition of this driver advantage speeds up the process of carcinogenesis therefore increasing the risk without necessarily leaving a sign on the mutation rate. Overall, most likely a combination of different factors are responsible for the progression to liver cancer in HCV-positive patients.

On the other hand, the absence on an increase in mutation rate in obese people, seems to indicate no role of mutations in increasing cancer risk among obese people. This needs further investigation and we have ongoing work testing an association between organ size and obesity.

# References


Song, Mingyang, Bert Vogelstein, Edward L Giovannucci, Walter C Willett, and Cristian Tomasetti (2018). "Cancer prevention: Molecular and epidemiologic consensus". In: *Science* 361.6409, pp. 1317–1318.

Cancer Research UK (2018). *Preventable cancers*. URL: www.cancerresearchuk.org/health-professional/cancer-statistics/risk/preventable-cancers (visited on 10/20/2018).

Health, US Department of, Human Services, et al. (2004). "The health consequences of smoking: a report of the Surgeon General". In:

Tomasetti, Cristian and Bert Vogelstein (2015). "Variation in cancer risk among tissues can be explained by the number of stem cell divisions". In: *Science* 347.6217, pp. 78–81.

Tomasetti, Cristian, Lu Li, and Bert Vogelstein (2017). "Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention". In: *Science* 355.6331, pp. 1330–1334.

Tomasetti, Cristian, Bert Vogelstein, and Giovanni Parmigiani (2013). "Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation". In: *Proceedings of the National Academy of Sciences* 110.6, pp. 1999–2004.

Tomasetti, Cristian, Luigi Marchionni, Martin A Nowak, Giovanni Parmigiani, and Bert Vogelstein (2015). "Only three driver gene mutations are required for the development of lung and colorectal cancers". In: *Proceedings of the National Academy of Sciences* 112.1, pp. 118–123.

Durrett, Richard and Stephen Moseley (2010). "Evolution of resistance and progression to disease during clonal expansion of cancer". In: *Theoretical population biology* 77.1, pp. 42–48.

Beaumont, Mark A, Wenyang Zhang, and David J Balding (2002). "Approximate Bayesian computation in population genetics". In: *Genetics* 162.4, pp. 2025–2035.

The Cancer Genome Atlas (2018). *TCGA*. URL: https://cancergenome.nih.gov/ (visited on 10/20/2018).

Pesch, Beate, Benjamin Kendzia, Per Gustavsson, Karl-Heinz Jöckel, Georg Johnen, Hermann Pohlabeln, Ann Olsson, Wolfgang Ahrens, Isabelle Mercedes Gross, Irene Brüske, et al. (2012). "Cigarette smoking and lung cancer–relative risk estimates for the major histological types from a pooled analysis of case–control studies". In: *International journal of cancer* 131.5, pp. 1210–1219.

Patel, Neel H, Kristopher M Attwood, Michael Hanzly, Terrance T Creighton, Diana C Mehedint, Thomas Schwaab, and Eric C Kauffman (2015). "Comparative analysis of smoking as a risk factor among renal cell carcinoma histological subtypes". In: *The Journal of urology* 194.3, pp. 640–646.

Hunt, Jay D, Olga L Van der Hel, Garnett P McMillan, Paolo Boffetta, and Paul Brennan (2005). "Renal cell carcinoma in relation to cigarette smoking: meta-analysis of 24 studies". In: *International journal of cancer* 114.1, pp. 101–108.

Maasland, Denise HE, Piet A van den Brandt, Bernd Kremer, R Alexandra Sandra Goldbohm, and Leo J Schouten (2014). "Alcohol consumption, cigarette smoking and the risk of subtypes of head-neck cancer: results from the Netherlands Cohort Study". In: *BMC cancer* 14.1, p. 187.

Freedman, Neal D, Debra T Silverman, Albert R Hollenbeck, Arthur Schatzkin, and Christian C Abnet (2011). "Association between smoking and risk of bladder cancer among men and women". In: *Jama* 306.7, pp. 737–745.

Calle, Eugenia E, Carmen Rodriguez, Kimberly Walker-Thurmond, and Michael J Thun (2003). "Overweight, obesity, and mortality from cancer in a prospectively studied cohort of US adults". In: *New England Journal of Medicine* 348.17, pp. 1625–1638.

Donato, F, A Tagger, U Gelatti, G Parrinello, P Boffetta, A Albertini, A Decarli, P Trevisi, ML Ribero, C Martelli, et al. (2002). "Alcohol and hepatocellular carcinoma: the effect of lifetime intake and hepatitis virus infections in men and women". In: *American journal of epidemiology* 155.4, pp. 323–331.

Stoffel, Elena, Bhramar Mukherjee, Victoria M Raymond, Nabihah Tayob, Fay Kastrinos, Jennifer Sparr, Fei Wang, Prathap Bandipalliam, Sapna Syngal, and Stephen B Gruber (2009). "Calculation of risk of colorectal and endometrial cancer among patients with Lynch syndrome". In: *Gastroenterology* 137.5, pp. 1621–1627.

Pfeifer, Gerd P, Mikhail F Denissenko, Magali Olivier, Natalia Tretyakova, Stephen S Hecht, and Pierre Hainaut (2002). "Tobacco smoke carcinogens,

DNA damage and p53 mutations in smoking-associated cancers". In: *oncogene* 21.48, p. 7435.

McGivern, David R and Stanley M Lemon (2011). "Virus-specific mechanisms of carcinogenesis in hepatitis C virus associated liver cancer". In: *Oncogene* 30.17, p. 1969.

# Chapter 4

# Statistical methods for analyzing liquid biopsy data

## 4.1 Introduction

Previous study suggests that mutations occurring during normal cell divides (R mutations) are responsible for two-thirds of the mutations in human cancers(Tomasetti, Li, and Vogelstein, 2017). As a result, many cancers arising from these unavoidable R mutations are not preventable. From public health perspective, this poses the challenge of reducing cancer deaths. It is conceivable that the best way to lower cancer mortality rates is primary prevention. HPV vaccine is one of the most successful primary prevention strategies that can significantly reduce the risk of developing HPV-associated cancers (National Cancer Institute, 2018a). While primary prevention is not feasible now for all cancer types, we still have secondary prevention, i.e. early detection and intervention.

Earlier detection is key to reducing cancer deaths. The majority of localized cancers can be cured by surgery alone, without any systemic therapy (Siegel,

Miller, and Jemal, 2017). Once distant metastasis has occurred, however, surgical excision is rarely curative. One major goal in cancer research is therefore the detection of cancers before they metastasize to distant sites. For many adult cancers, it takes 20 to 30 years for incipient neoplastic lesions to progress to late-stage disease (Vogelstein et al., 2013; Jones et al., 2008; Yachida et al., 2012). Only in the past few years of this long process do neoplastic cells appear to successfully seed and give rise to metastatic lesions (Vogelstein et al., 2013; Jones et al., 2008; Yachida et al., 2012; Vogelstein and Kinzler, 2015). Thus, there is a wide window of opportunity to detect cancers before the onset of metastasis. Even when metastasis has initiated but is not yet evident radiologically, cancers can be cured in up to 50% of cases with systemic therapies, such as cytotoxic drugs and immunotherapy (Bozic et al., 2013; Semrad et al., 2015; Moertel et al., 1995; Huang et al., 2017). Once large, metastatic tumors are formed, however, current therapies are rarely effective (Bozic et al., 2013; Semrad et al., 2015; Moertel et al., 1995; Huang et al., 2017).

The importance of cancer early detection can also be seen from the 5-year survival rate. For instance, as is shown in Figure 4.1, the 5-year survival rate for localized lung cancer is 56.3%. However, when the disease has spread to regional lymph nodes or metastasized, the survival rate drops to 29.7% or 4.7% respectively. On the other hand, when looking at the percentage of cases at diagnosis, only 16% are localized disease (National Cancer Institute, 2018b). These facts indicate that we can potentially reduce the cancer mortality rate significantly by detecting more cancers at earlier stage.

We describe here a new blood test, called CancerSEEK, that can detect eight

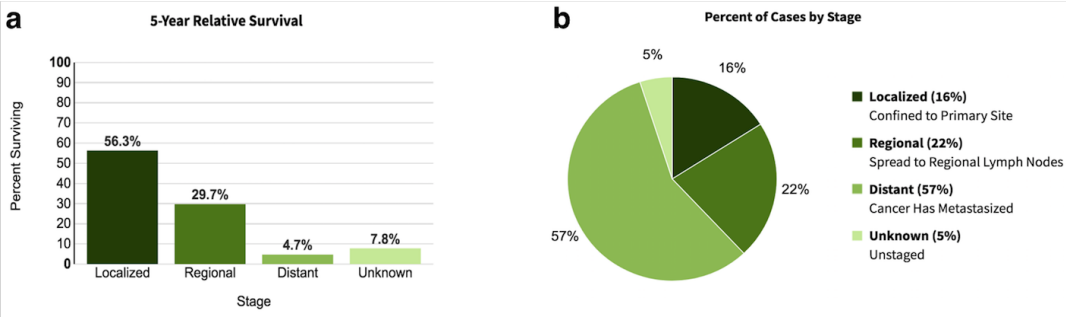


**Figure 4.1: Percent of Cases & 5-Year Relative Survival by Stage at Diagnosis: Lung and Bronchus Cancer** SEER 18 2008 - 2014, All Races, Both Sexes by SEER Summary Stage 2000 (a) 5-year survival by Stage. Localized disease has much higher survival rate than regional and metastasized cancers. (b) Percent of cases by stage. Although localized disease has much higher survival rate, it only accounts for 16% of all lung cancer cases. Source: Surveillance, Epidemiology, and End Results Program (SEER)

common cancer types (ovary, liver, stomach, pancreas, esophagus, colorectum, lung, and breast) through assessment of the levels of circulating proteins and mutations in cell-free DNA, and has the capacity not only to identify the presence of relatively early cancers but also to localize the organ of origin of these cancers.

## 4.2 Statistical Challenges

### 4.2.1 Experimental design

We began by designing a polymerase chain reaction (PCR)-based assay that could simultaneously assess multiple regions of driver genes that are commonly mutated in a variety of cancer types. A 61-amplicon panel, with each amplicon querying an average of 33 base pairs (bp) within one of 16 genes, was used in the plasma DNA-based tests. As shown in Figure 4.2, this

panel would theoretically detect 41% (liver) to 95% (pancreas) of the cancers in the Catalog of Somatic Mutations in Cancer (COSMIC) data set (Forbes et al., 2016). In practice, the panel performed considerably better, detecting at least one mutation in 82%, two mutations in 47%, and more than two mutations in 8% of the 805 cancers evaluated in our study (Figure 4.2, colored dots). We were able to detect a larger fraction of tumors than predicted by the COSMIC data set because the PCR-based sequencing assay we used was more sensitive for detecting mutations than conventional genome-wide sequencing. On the basis of this analysis of the DNA from primary tumors, the predicted maximum detection capability of circulating tumor DNA (ctDNA) in our study varied by tumor type, ranging from 60% for liver cancers to 100% for ovarian cancers (Figure 4.2).

Armed with this small but robust panel of amplicons, we developed two approaches that enabled the detection of the rare mutations expected to be present in plasma ctDNA. First, we used multiplex-PCR to directly and uniquely label each original template molecule with a DNA barcode. This design minimizes the errors inherent to massively parallel sequencing (Kinde et al., 2011) and makes efficient use of the small amount of cell-free DNA present in plasma. Additionally, we divided the total amount of DNA recovered from plasma into multiple aliquots and performed independent assays on each replicate. In effect, this decreases the number of DNA molecules per well; however, it increases the fraction of each mutant molecule per well, making the mutants easier to detect. Because the sensitivity of detection is

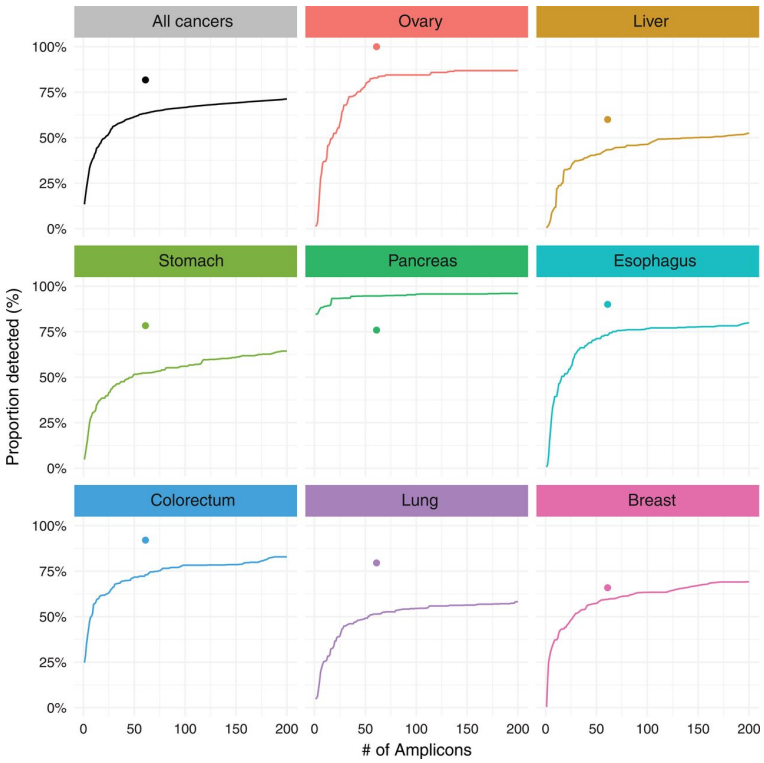


**Figure 4.2: Development of a PCR-based assay to identify tumor-specific mutations in plasma samples** Colored curves indicate the proportion of cancers of the eight types evaluated in this study that can be detected with an increasing number of short ($<$ 40 bp) amplicons. The sensitivity of detection increases with the number of amplicons but plateaus at about 60 amplicons. Colored dots indicate the fraction of cancers detected by using the 61-amplicon panel used in 805 cancers evaluated in our study, which averaged 82%. Publicly available sequencing data were obtained from the COSMIC repository.

often limited by the fraction of mutant alleles in each replicate, this partitioning strategy allowed us to increase the signal-to-noise ratio and identify mutations present at lower prevalence than possible if all of the plasma DNA was evaluated at once. As a result, each individual sample was tested in 6 independent wells, with roughly the same amount of DNA per well.

The second component of CancerSEEK is based on protein biomarkers. Previous studies have demonstrated that a major fraction of early-stage tumors do not release detectable amounts of ctDNA, even when extremely sensitive techniques are used to identify them (Bettegowda et al., 2014; Cohen et al., 2017). Many proteins potentially useful for early detection and diagnosis of cancer have been described in the literature (Liotta et al., 2003; Patz Jr et al., 2007; Wang et al., 2016). We searched this literature to find proteins that had previously been shown to detect at least one of the eight cancer types described above with sensitivities >10% and specificities >99%. We found that 39 of these proteins could be reproducibly evaluated through a single immunoassay platform, and we then used this platform to assay all plasma samples.

### 4.2.2 Challenges in mutation analysis

We were confronted by three competing challenges when developing the algorithm for the mutation analysis. First, although a more sensitive approach, as described above, was applied in the sequencing to detect rare mutations, the PCR errors are unavoidable. It arises from different error-generating processes, including but not limited to errors introduced during

PCR amplification to prepare the libraries for capturing or for inverse PCR, errors introduced by other enzymatic steps, particularly if the enzymes are impure and contaminated, errors introduced during the shearing process required to generate small fragments for sequencing, etc.(Kinde et al., 2011) One critical challenge for the mutation analysis is to separate technical artifacts from true mutations. Second, for unknown reasons, some positions have higher background error rates than others. Normalization step is necessary to adjust for these different error rates before any analysis. Third, the algorithm should combine the information from 6 independent measurements of each mutation.

### 4.2.3 Challenges in protein analysis

We also faced several challenges in protein analysis. First of all, the analysis should account for the variations in upper and lower limits of detection across different experiments. Secondly, valid protein biomarkers should be selected from the original 39 proteins. Thirdly, the algorithm should incorporate information from both mutation and protein analysis.

## 4.3 Algorithms for CancerSEEK analysis

### 4.3.1 Mutation Analysis

#### 4.3.1.1 Normalization

All mutations that did not have more than 1 supermutant (SM) in at least one well were excluded from the analysis. The mutant allele frequency (MAF), defined as the ratio between the total number of supermutants in

each well from that sample and the total number of UIDs in the same well from that sample, was first normalized based on the observed MAFs for each mutation in a set of normal controls comprising the normal plasmas in the training set plus a set of 256 white blood cell samples (WBCs) from unrelated healthy individuals. All MAFs with <100 UIDs were set to zero. This normalization was performed by first calculating the average MAF ($\overline{MAF_j^{Ref}}$) for each mutation $j = 1, \cdots, n$, found among the normal controls. Using the 25th percentile of the distribution generated by these averages as the reference value $q \triangleq Q(\frac{1}{4}, \overline{MAF_\bullet^{Ref}})$, each MAF was normalized multiplying it by the ratio $\frac{q}{\overline{MAF_j^{Ref}}}$. For example, if the observed average MAF of a mutation in a set of controls was 10 times higher than $q$, then each MAF for that mutation was multiplied by $\frac{1}{10}$. Assuming total number of samples in the test set $m$, total number of mutations observed in the reference set $n$,

$$MAF_{ij}^{adj} = MAF_{ij}^{obs} \cdot \frac{q}{\overline{MAF_j^{Ref}}}, i = 1, \cdots, m; j = 1, \cdots, n$$

If a mutation in a test sample was not observed in any normal control, it was not normalized, i.e. observed MAF was used for the analysis. In this way, uniformly high MAFs in certain mutations, if also present in reference set for any reason, will be reduced. This helps to increase specificity. Standard normalization, i.e. subtracting the mean and dividing by the standard deviation, did not perform as well according to the cross-validation results.

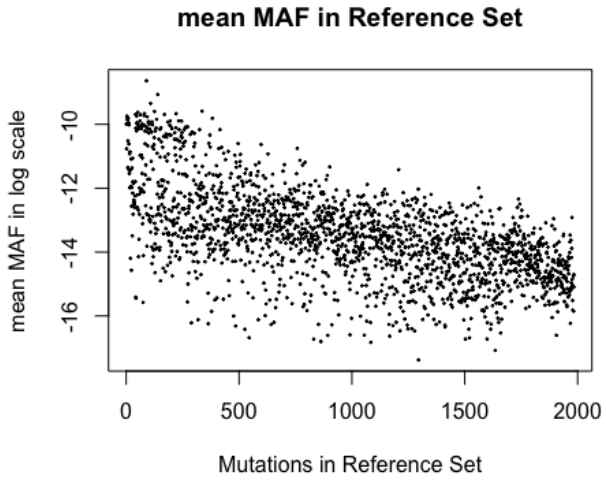


**Figure 4.3: Distributions of average MAF in the reference set** Average MAF by mutation is calculated for all mutations in the reference set using all qualified wells. MAF is shown in log scale in the plot.

#### 4.3.1.2 Reference distributions and p-values

Following this mutation-specific normalization, the UID range was split in 10 intervals ($< 1,000, 1000 - 2000, \cdots, 8000 - 9000, > 9000$). Groups with more UIDs tend to have lower MAFs, as is shown in Figure 4.4. This group-specific distributions enable us to be more sensitive on sample classifications. Depending on the number of UIDs, the MAF of each mutation in each well was compared to two reference distributions of MAFs built from samples in the corresponding UID range: 1) a distribution built from all the normal control plasmas in the training set plus a set of 256 WBCs from unrelated, healthy individuals, denoted as $D_k^N, k = 1, \cdots, 10$; and 2) a distribution built from the plasma samples from cancer patients in the training set, denoted as $D_k^C, k = 1, \cdots, 10$. The cancer training set included only those in which the same mutation was present in the plasma and in the corresponding primary

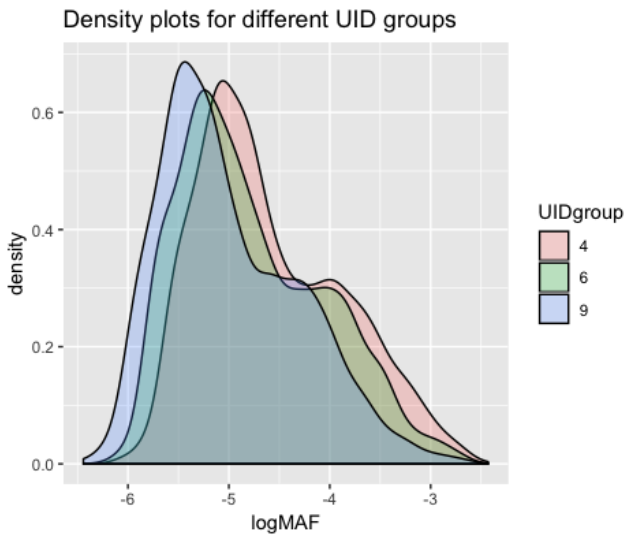


**Figure 4.4: Density plot of log MAF in different UID groups** Distributions of normal control MAFs in log scale are shown for UID group 4, 6, and 9 respectively. The shapes are in general similar, but the means are somewhat different across UID groups.

tumor, with an $MAF > 5\%$ in the tumor. Corresponding p-values, $p^N$ and $p^C$, were thus obtained. The reference distributions for both the normal and cancer samples were built independently, from the training sets, in each round and each iteration of 10-fold cross-validation, i.e., 90% of the samples in each iteration were used for training and 10% of the samples were used for testing. Specifically, each well of every mutation had 2 p-values,

$$p_{ij}^N = Pr(X \geq MAF_{ij}^{adj} | X \sim D_k^N, UID_{ij} \in UIDgroup_k)$$

$$p_{ij}^C = Pr(X \leq MAF_{ij}^{adj} | X \sim D_k^C, UID_{ij} \in UIDgroup_k)$$

#### 4.3.1.3 Log ratios and Omega score

For each well, the log ratio of these two pvalues, $p_{ij}^C / p_{ij}^N$ was then calculated, and the minimum and maximum of these log ratios across the six wells for a mutation were eliminated so that the results would be less sensitive to outliers. We considered the log ratio of the p-values rather than the standard log-likelihood ratio because the relatively low number of data points available did not allow a robust estimation of the densities of the MAF distributions (particularly for $p^C$). An "Omega" score was then determined according to the following formula:

$$\Omega = \sum w_i \log \frac{p_{ij}^C}{p_{ij}^N}$$

where $w_i$ is the number of UIDs in well $i$ divided by the total number of UIDs for that mutation in the four wells that were included in the analysis (the two outlying wells were excluded, as noted above). We weighted the log ratio of p-values so that those wells containing more template molecules would have a greater impact on the final statistic (the omega score). The rationale for this weighting was that the larger the number of template molecules in a well, the more confidence in the result.

To further illustrate how the omega score is obtained, a specific example of its calculation is provided in Table 4.1. Consider the $KRASp.G12S, c.34G > A$ mutation found in sample $IND I256 PLS1$. By eliminating the minimum and maximum values of those ratios, and applying the above formula for omega,

we obtain the omega score for that mutation:

$$\Omega = \frac{3755}{11393}\log 94243 + \frac{2198}{11393}\log 21716 + \frac{2013}{11393}\log 110752 + \frac{3427}{11393}\log 12680$$

$$= 10.60$$

| | Well 1 | Well 2 | Well 3 | Well 4 | Well 5 | Well 6 |
|---|---|---|---|---|---|---|
| SM | 161 | 78 | 99 | 84 | 177 | 117 |
| UID | 3755 | 2198 | 2966 | 2013 | 3694 | 3427 |
| MAF | 0.043 | 0.035 | 0.033 | 0.042 | 0.048 | 0.034 |
| adj MAF | 0.0057 | 0.0047 | 0.0044 | 0.0056 | 0.0064 | 0.004 |
| $p^N$ | 1.06e-06 | 5.70e-06 | 1.02e-05 | 1.03e-06 | 3.09e-07 | 8.83e-06 |
| $p^C$ | 0.100 | 0.124 | 0.128 | 0.114 | 0.094 | 0.112 |
| ratio($\frac{p^C}{p^N}$) | 94243 | 21716 | 12510 | 110752 | 305090 | 12680 |

**Table 4.1: Example: Mutation analysis**
This example illustrate the analysis using $KRASp.G12S, c.34G > A$ mutation in sample INDI 256 PLS 1. The final omega score for this mutation is 10.6

When a mutation identified in a plasma sample had $\Omega > 1$, and was not identified in the primary tumor of the patient, we evaluated DNA from white blood cells (WBCs) of the same patient whenever WBCs were available (23% of the cancer patients). WBC DNA was tested with the same 61-amplicon panel to ensure that the plasma mutation was not a result of Clonal Hematopoiesis of Indeterminate Potential (Jaiswal et al., 2014). WBCs from the normal individuals were evaluated identically whenever a mutation with $\Omega > 1$ was found in the plasma. Any mutation that was identified in the WBCs as well as in the plasma was excluded from the analysis. The requirement for exclusion was that the ratio between the max MAF in the plasma and the max MAF in the WBC was less than 100. The mutation with the greatest $\Omega$ score in each patient or normal control was then deemed the "top mutation".

### 4.3.2 Protein Analysis

#### 4.3.2.1 Protein's normalization and transformation

To account for the variations in the lower and upper limits of detection across different experiments, we set all values smaller than m, defined as the maximum among all lower limits of detection for a given protein among all experiments, equal to m. By symmetry, we set all values larger than M, defined as the minimum among all upper limits of detection for that protein across all experiments, equal to M. To be conservative, a further transformation was applied to the proteins levels. Specifically, if a protein's concentration in the sample of interest was lower than the 95th percentile of the concentration found for that same protein among the normal samples in the training set, then the protein's concentration was set equal to zero; otherwise its original concentration value was used. For the $\Omega$ score, the same threshold transformation was used but with a constant threshold equal to 0, because $\Omega > 0$ indicates an MAF that is more likely to originate from a cancer than from normal tissue.

#### 4.3.2.2 Combining proteins and mutations

The omega score was used as a feature in logistic regression (LR). The other 8 features used in LR were the concentrations of the following 8 proteins, selected from the original 39 proteins via a straightforward optimization: CA125, CA19-9, CEA, HGF, MPO, OPN, PRL, TIMP-1. The optimization first eliminated any protein that, according to Mann-Whitney-Wilcoxon test, had higher median values in normal than in cancer samples, eliminating 13

proteins and leaving 26 proteins to be evaluated. This was followed by a forward selection based on the importance of each feature, as evaluated by the decrease in accuracy of the same logistic regression when that protein alone was dropped from the remaining 26 protein features. Ten rounds of 10-fold cross-validations were performed.

### 4.3.3 Concordance check

For determining the concordance between mutations identified in the plasma with those identified in primary tumors, we only considered the 153 cases in which a mutation could be identified with high confidence in the plasma ($\Omega > 3$) and in which the primary tumor contained any mutation that was present at a mutant allele fraction of $> 5\%$. This approach allowed us to avoid scoring tumors that had low neoplastic contents (Makohon-Moore et al., 2017).

## 4.4 Results

We used CancerSEEK to study 1005 patients who had been diagnosed with stage I to III cancers of the ovary, liver, stomach, pancreas, esophagus, colorectum, lung, or breast. No patient received neo-adjuvant chemotherapy before blood sample collection, and none had evident distant metastasis at the time of study entry. The median age at diagnosis was 64 (range 22 to 93). The eight cancer types were chosen because they are common in western populations and because no blood-based tests for their earlier detection are in common clinical use. The most common stage at presentation was American

Joint Commission on Cancer (AJCC) stage II, accounting for 49% of patients, with the remaining patients harboring stage I (20%) or stage III (31%) disease. The healthy control cohort consisted of 812 individuals of median age 55 (range 17 to 88) with no known history of cancer, high-grade dysplasia, autoimmune disease, or chronic kidney disease.

The mean sensitivities and specificities were determined by 10 iterations of 10-fold cross-validations. The receiver operating characteristic (ROC) curves for the entire cohort of cancer patients and controls in one representative iteration is shown in Figure 4.5.

The median sensitivity of CancerSEEK among the eight cancer types evaluated was 70% and ranged from 98% in ovarian cancers to 33% in breast cancers (Figure 4.5C). At this sensitivity, the specificity was >99%; only 7 of the 812 individuals without known cancers scored positive. We could not be certain that the few false positive-testing individuals identified among the healthy cohort did not actually have an as-yet undetected cancer, but classifying them as false positives provided the most conservative approach to classification and interpretation of the data.

One of the most important attributes of a screening test is the ability to detect cancers at relatively early stages. The median sensitivity of CancerSEEK was 73% for the most common stage evaluated (stage II), similar (78%) for stage III cancers, and lower (43%) for stage I cancers (Figure 4B). The sensitivity for the earliest-stage cancers (stage I) was highest for liver cancer (100%) and lowest for esophageal cancer (20%).

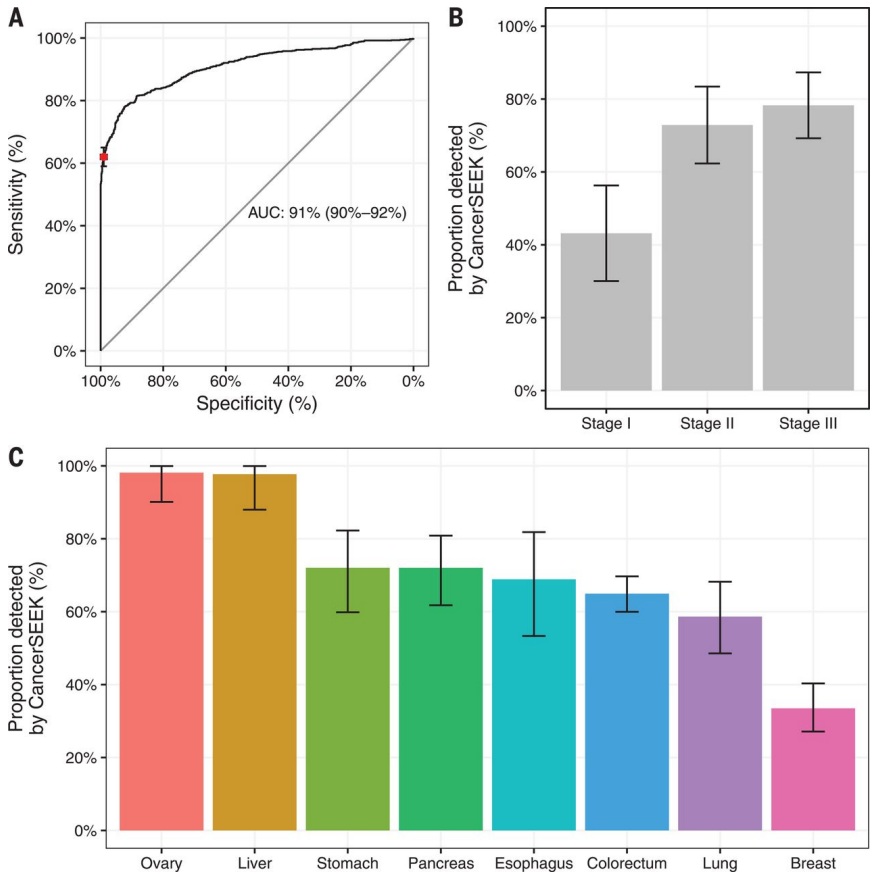


**Figure 4.5: Performance of CancerSEEK** (A) ROC curve for CancerSEEK. The red point on the curve indicates the test's average performance (62%) at >99% specificity. Error bars represent 95% confidence intervals for sensitivity and specificity at this particular point. The median performance among the eight cancer types assessed was 70%. (B) Sensitivity of CancerSEEK by stage. Bars represent the median sensitivity of the eight cancer types, and error bars represent standard errors of the median. (C) Sensitivity of CancerSEEK by tumor type. Error bars represent 95% confidence intervals.

## 4.5 Discussions

We have designed a multi-analyte blood test that can detect the presence of eight common solid tumor types. The advantage of combining completely different agents, with distinct mechanisms of action, is widely recognized in therapeutics (Organization and Organization), 2010; Organization et al., 2016; Benson et al., 2017) but has not been routinely applied to diagnostics. We combined protein biomarkers with genetic biomarkers to increase sensitivity without substantially decreasing specificity. Other cancer biomarkers-such as metabolites, mRNA transcripts, miRNAs, or methylated DNA sequences-could be similarly combined to increase sensitivity and localization of cancer site. Such multi-analyte tests are not meant to replace other non-blood-based screening tests, such as those for breast or colorectal cancers, but to provide additional information that could help identify those patients most likely to harbor a malignancy.

Several limitations of our study should be acknowledged. First, the patient cohort in our study was composed of individuals with known cancers, most diagnosed on the basis of symptoms of disease. Although none of our patients had clinically evident metastatic disease at the time of study entry, most individuals in a true screening setting would have less advanced disease, and the sensitivity of detection is likely to be less than reported here. Second, our controls were limited to healthy individuals, whereas in a true cancer screening setting, some individuals might have inflammatory or other diseases, which could result in a greater proportion of false-positive results than observed in our study. Third, although multiple-fold cross-validation is

a powerful and widely used technique for demonstrating robust sensitivity and specificity on a cohort of this study's scale, we were not able to use a completely independent set of cases for testing, which would have been optimal. Last, the proportion of cancers of each type in our cohort was purposefully not representative of those in the United States as a whole because we wanted to evaluate at least 50 examples of each cancer type with the resources available to us. When weighted for actual incidence in the United States, we estimate the sensitivity of CancerSEEK to be 55% among all eight cancer types. This weighting would not affect the high sensitivities of CancerSEEK (69 to 98%) to detect five cancer types (ovary, liver, stomach, pancreas, and esophagus) for which there are no screening tests available for average-risk individuals.

Our study lays the conceptual and practical foundation for a single, multianalyte blood test for cancers of many types. We estimate the cost of the test to be less than $500, which is comparable or lower than other screening tests for single cancers, such as colonoscopy. The eight cancer types studied here account for 360,000 (60%) of the estimated cancer deaths in the United States in 2017, and their earlier detection could conceivably reduce deaths from these diseases. To actually establish the clinical utility of CancerSEEK and to demonstrate that it can save lives, prospective studies of all incident cancer types in a large population will be required.

# References


Tomasetti, Cristian, Lu Li, and Bert Vogelstein (2017). "Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention". In: *Science* 355.6331, pp. 1330–1334.

National Cancer Institute (2018a). *Human Papillomavirus (HPV) Vaccines*. URL: https://www.cancer.gov/about-cancer/causes-prevention/risk/infectious-agents/hpv-vaccine-fact-sheet (visited on 10/20/2018).

Siegel, Rebecca L, Kimberly D Miller, and Ahmedin Jemal (2017). "Cancer statistics, 2017". In: *CA: a cancer journal for clinicians* 67.1, pp. 7–30.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler (2013). "Cancer genome landscapes". In: *science* 339.6127, pp. 1546–1558.

Jones, Siân, Wei-dong Chen, Giovanni Parmigiani, Frank Diehl, Niko Beerenwinkel, Tibor Antal, Arne Traulsen, Martin A Nowak, Christopher Siegel, Victor E Velculescu, et al. (2008). "Comparative lesion sequencing provides insights into tumor evolution". In: *Proceedings of the National Academy of Sciences* 105.11, pp. 4283–4288.

Yachida, Shinichi, Catherine White, Yoshiki Naito, Yi Zhong, Jacqueline A Brosnan, Anne M Macgregor-Das, Richard A Morgan, Tyler Saunders, Daniel Laheru, Joseph M Herman, et al. (2012). "Clinical significance of the genetic landscape of pancreatic cancer and implications for identification of potential long term survivors". In: *Clinical Cancer Research*, clincanres–1215.

Vogelstein, Bert and Kenneth W Kinzler (2015). "The path to cancer-three strikes and you're out". In: *N Engl J Med* 373.20, pp. 1895–1898.

Bozic, Ivana, Johannes G Reiter, Benjamin Allen, Tibor Antal, Krishnendu Chatterjee, Preya Shah, Yo Sup Moon, Amin Yaqubie, Nicole Kelly, Dung T Le, et al. (2013). "Evolutionary dynamics of cancer in response to targeted combination therapy". In: *elife* 2, e00747.

Semrad, Thomas J, Ana Rodriguez Fahrni, I-Yeh Gong, and Vijay P Khatri (2015). "Integrating chemotherapy into the management of oligometastatic colorectal cancer: Evidence-based approach using clinical trial findings". In: *Annals of surgical oncology* 22.3, pp. 855–862.

Moertel, Charles G, Thomas R Fleming, John S Macdonald, Daniel G Haller, John A Laurie, Catherine M Tangen, James S Ungerleider, William A Emerson, Douglass C Tormey, John H Glick, et al. (1995). "Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: a final report". In: *Annals of internal medicine* 122.5, pp. 321–326.

Huang, Alexander C, Michael A Postow, Robert J Orlowski, Rosemarie Mick, Bertram Bengsch, Sasikanth Manne, Wei Xu, Shannon Harmon, Josephine R Giles, Brandon Wenz, et al. (2017). "T-cell invigoration to tumour burden ratio associated with anti-PD-1 response". In: *Nature* 545.7652, p. 60.

National Cancer Institute (2018b). *Percent of Cases & 5-Year Relative Survival by Stage at Diagnosis: Lung and Bronchus Cancer*. URL: https://seer.cancer.gov/statfacts/html/lungb.html (visited on 10/20/2018).

Forbes, Simon A, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, et al. (2016). "COSMIC: somatic cancer genetics at high-resolution". In: *Nucleic acids research* 45.D1, pp. D777–D783.

Kinde, Isaac, Jian Wu, Nick Papadopoulos, Kenneth W Kinzler, and Bert Vogelstein (2011). "Detection and quantification of rare mutations with massively parallel sequencing". In: *Proceedings of the National Academy of Sciences* 108.23, pp. 9530–9535.

Bettegowda, Chetan, Mark Sausen, Rebecca J Leary, Isaac Kinde, Yuxuan Wang, Nishant Agrawal, Bjarne R Bartlett, Hao Wang, Brandon Luber, Rhoda M Alani, et al. (2014). "Detection of circulating tumor DNA in early- and late-stage human malignancies". In: *Science translational medicine* 6.224, 224ra24–224ra24.

Cohen, Joshua D, Ammar A Javed, Christopher Thoburn, Fay Wong, Jeanne Tie, Peter Gibbs, C Max Schmidt, Michele T Yip-Schneider, Peter J Allen, Mark Schattner, et al. (2017). "Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers". In: *Proceedings of the National Academy of Sciences* 114.38, pp. 10202–10207.

Liotta, LA et al. (2003). "The promise of proteomics." In: *Clinical advances in hematology & oncology: H&O* 1.8, pp. 460–462.

Patz Jr, Edward F, Michael J Campa, Elizabeth B Gottlin, Irina Kusmartseva, Xiang Rong Guan, and James E Herndon (2007). "Panel of serum biomarkers for the diagnosis of lung cancer". In: *Journal of Clinical Oncology* 25.35, pp. 5578–5583.

Wang, Hui, Tujin Shi, Wei-Jun Qian, Tao Liu, Jacob Kagan, Sudhir Srivastava, Richard D Smith, Karin D Rodland, and David G Camp (2016). "The clinical impact of recent advances in LC-MS for cancer biomarker discovery and verification". In: *Expert review of proteomics* 13.1, pp. 99–114.

Jaiswal, Siddhartha, Pierre Fontanillas, Jason Flannick, Alisa Manning, Peter V Grauman, Brenton G Mar, R Coleman Lindsley, Craig H Mermel, Noel Burtt, Alejandro Chavez, et al. (2014). "Age-related clonal hematopoiesis associated with adverse outcomes". In: *New England Journal of Medicine* 371.26, pp. 2488–2498.

Makohon-Moore, Alvin P, Ming Zhang, Johannes G Reiter, Ivana Bozic, Benjamin Allen, Deepanjan Kundu, Krishnendu Chatterjee, Fay Wong, Yuchen Jiao, Zachary A Kohutek, et al. (2017). "Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer". In: *Nature genetics* 49.3, p. 358.

Organization, World Health and Stop TB Initiative (World Health Organization) (2010). *Treatment of tuberculosis: guidelines*. World Health Organization.

Organization, World Health et al. (2016). *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach*. World Health Organization.

Benson, Al B, Alan P Venook, Lynette Cederquist, Emily Chan, Yi-Jen Chen, Harry S Cooper, Dustin Deming, Paul F Engstrom, Peter C Enzinger, Alessandro Fichera, et al. (2017). "Colon cancer, version 1.2017, NCCN clinical practice guidelines in oncology". In: *Journal of the National Comprehensive Cancer Network* 15.3, pp. 370–398.

# Chapter 5

# Conclusion and Future work

In this dissertation, we proposed novel approaches to address the importance of random mutations in cancer causation, and developed methods for analyzing liquid biopsy data for cancer early detection.

Specifically, in Chapter 2, we used genome sequencing and epidemiological data to determine the proportions of cancer-causing mutations that result from inherited (H), environmental (E), and replicative factors (R) (Tomasetti, Li, and Vogelstein, 2017). Our results highlight the prominent role of R mutations in cancer etiology and conclude that a substantial amount of cancer-causing mutations are actually due to the random replicative errors. These findings are consistent with the epidemiological evidence on the fraction of preventable cancers, and in fact, provide a well-defined, molecular explanation for the large and apparently unpreventable component of cancer risk that has long puzzled epidemiologists.

The above result consider exclusively mutations. However, independently of whether these mutations are contributed by E, H, or R, a fundamental but completely unanswered question remains: how much of cancer is due to

mutations? In Chapter 3, we sought to estimate the proportions of increased cancer risk that can be explained by the increased mutation rate. Our results suggest that mutations account for an extremely large proportion of the higher risk due to smoking and microsatellite instability (MSI), a relatively large proportion in Hepatitis C infections, but almost nothing in obesity. Future work is needed to understand what is the mechanism through which obesity works in increasing cancer risk.

These two studies point to the fact that not all cancers are preventable due to the unavoidable R mutations. Therefore, in Chapter 4, we developed statistical methods for CancerSEEK, a multi-analyte blood test for cancer early detection (Cohen et al., 2018). With 99% specificity, our sensitivities ranged from 69% to 98% for the detection of five cancer types (ovary, liver, stomach, pancreas, and esophagus) for which there are no screening tests available for average-risk individuals.

While the majority of cancer research is focused on curing late stage diseases, the data from CancerSEEK study presents promising result on detecting early cancer cases using a blood-based test. However, due to the low disease prevalence, it should be evaluated using large clinical trials in a real-word setting before being approved as a clinically useful screening test. In addition, the sensitivities vary across different cancer types and different stages. Looking at the sensitivities, it is clear that we have a large room for improvement, like for breast cancer and stage I disease. It is also worthwhile to mention that even though the test has a very high specificity (>99%), it is likely that we will still have a substantial amount of false positives when applying the screening

test in the general population due to the low disease prevalence.

Although in CancerSEEK we analyzed mutations from ctDNA and protein biomarkers, there are also other panels that can potentially increase the sensitivity of the test, including but not limited to, assay for aneuploidy (Adalsteinsson et al., 2017), tissue-specific DNA methylation patterns (Snyder et al., 2016). Combining multiple blood-based analytes to achieve optimal sensitivities, while maintaining cost-effective and high specificity as a screening test, represents, in my opinion, a fundamental research direction of this field in the future.

# References

Tomasetti, Cristian, Lu Li, and Bert Vogelstein (2017). "Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention". In: *Science* 355.6331, pp. 1330–1334.

Cohen, Joshua D, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A Javed, Fay Wong, Austin Mattox, et al. (2018). "Detection and localization of surgically resectable cancers with a multi-analyte blood test". In: *Science*, eaar3247.

Adalsteinsson, Viktor A, Gavin Ha, Samuel S Freeman, Atish D Choudhury, Daniel G Stover, Heather A Parsons, Gregory Gydush, Sarah C Reed, Denisse Rotem, Justin Rhoades, et al. (2017). "Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors". In: *Nature communications* 8.1, p. 1324.

Snyder, Matthew W, Martin Kircher, Andrew J Hill, Riza M Daza, and Jay Shendure (2016). "Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin". In: *Cell* 164.1, pp. 57–68.

# Lu Li

615 N Wolfe Street, E3031, Baltimore MD 21205
+1 443-510-8841 • lli48@jhu.edu

## Education

**Johns Hopkins University, Bloomberg School of Public Health** **Baltimore, MD USA**
***PhD in Biostatistics*** *Anticipated Nov 2018*
GPA: 4.0/4.0
Advisor: Dr. Cristian Tomasetti
Awarded Best Performance in PhD Qualifying Exam

**Johns Hopkins University** **Baltimore, MD USA**
*Master of Engineering in Financial Mathematics* *December 2013*
GPA: 3.9/4.0

**Wuhan University, Institute for Advanced Study** **Wuhan, China**
*Bachelor of Science in Mathematics, Bachelor of Arts in Economics* *June 2012*
GPA: 3.8/4.0

## PhD Research Highlights

- Develop statistical algorithms to analyze genetic data for cancer early detection and risk prediction
  - The algorithms for sample classification and tissue localization are the statistical methods part of CancerSEEK, a blood test for cancer early detection that was first published on Science in 2018, and have received broad media coverage by CNN, BBC, Forbes, etc.
- Use mathematical modeling to explain the role of genetic mutations in cancer etiology
  - This is a major part of the research on the "Bad Luck" Theory, which explains the role of random mutations in cancer etiology. This result published on Science in March 2017 has received broad media coverage by CNN, Washington Post, NPR, etc. and was also highlighted in NIH director blog.
- Make major contributions to multiple publications on high impact journals, such as **Science**, **Science Translational Medicine**, and **PNAS**

## Industry Experience

**Takeda Pharmaceutical Company** **Cambridge, MA USA**
*Reseach Intern - Statistics* *May - August 2018*
- Performed statistical research on Multiple Comparison Procedures and Modeling (MCP-Mod) in Phase II Dose-Finding Studies
- Conducted extensive simulations to evaluate the performance of MCP-Mod on univariate binary endpoint under different design assumptions
- Developed methods to extend the original MCP-Mod to co-primary binary endpoints
- Generated correlated binary responses and ran simulations to evaluate the new method
- Built an R Shiny app to visualize the dose-response models

**Towers Watson Consulting (now Willis Towers Watson)** **Beijing, China**
*Consultant Intern, Benefits Group* *June - August 2013*
- Performed research on data and policies of Chinese Social Insurance System and gave presentations
- Validated and analyzed data collected from clients
- Provided assistance in writing consulting report and prepared documents for presentations

**Deloitte Touche Tohmatsu Limited** **Beijing, China**
*Consultant Intern* *July – August 2010*
- Designed charts independently for the first two quarters' financial statements
- Interviewed managers at various levels to communicate client company operations
- Prepared status reports for further analysis and recommendations for the project

## Research Experiences

**Statistical Analysis of Genetic Data in Cancer Early Detection** **Baltimore, MD USA**
*Ludwig Center, Johns Hopkins School of Medicine* *June 2015 - present*
- Conduct simulations to optimize experimental designs of massively parallel sequencing
- Develop and implement algorithms based on permutation test, empirical Bayes estimation, probability distributions, and machine learning to classify ctDNA status, a main biomarker for cancer early detection
- Results have been published on high impact journals such as Science, Science Translational Medicine, GUT, etc.

**Statistical Methods on Cancer Etiology** **Baltimore, MD USA**
*Under supervision of Dr. Cristian Tomasetti* *June 2015 - present*
- Review major epidemiological studies on over 15 major cancer risk factors and summarizing main results
- Use mathematical modeling and statistical analysis to calculate the proportions of mutations that are inherited and induced by environmental factors respectively
- Develop statistical method to assess the impact of genetic mutations on cancer etiology
- First result has been published on Science in 2017

## Talks

**2019**: Statistical and Mathematical Approaches to Cancer Etiology, JSM (invited)

## Selected Publications


[1] Joshua D Cohen, **Lu Li**, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A Javed, Fay Wong, Austin Mattox, et al. "Detection and localization of surgically resectable cancers with a multi-analyte blood test". *Science* (2018). Developed statistical methods and analyzed data.

[2] Simeon U Springer, Chung-Hsin Chen, Maria Del Carmen Rodriguez Pena, **Lu Li**, Christopher Douville, Yuxuan Wang, Joshua David Cohen, Diana Taheri, Natalie Silliman, Joy Schaefer, et al. "Non-invasive detection of urothelial cancer through the analysis of driver gene mutations and aneuploidy". *eLife* 7 (2018). Developed algorithms for mutation analysis and interpreted results.

[3] Jeanne Tie, Joshua D Cohen, Yuxuan Wang, **Lu Li**, Michael Christie, Koen Simons, Hany Elsaleh, Suzanne Kosmider, Rachel Wong, Desmond Yip, et al. "Serial circulating tumour DNA analysis during multimodality treatment of locally advanced rectal cancer: a prospective biomarker study". *Gut* (2018). Developed the algorithm for classifying ctDNA status, analyzed and interpreted the data.

[4] Yuxuan Wang, **Lu Li**, Christopher Douville, Joshua D Cohen, Ting-Tai Yen, Isaac Kinde, Karin Sundfelt, Susanne K Kjær, Ralph H Hruban, Ie-Ming Shih, et al. "Evaluation of liquid from the Papanicolaou test and other liquid biopsies for the detection of endometrial and ovarian cancers". *Science translational medicine* 10.433 (2018). Developed algorithms for statistical analysis and analyzed data.

[5] Cristian Tomasetti, **Lu Li**, and Bert Vogelstein. "Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention". *Science* 355.6331 (2017). Obtained the estimates in tables S5 and S6.

[6] Jeanne Tie, Yuxuan Wang, Cristian Tomasetti, **Lu Li**, Simeon Springer, Isaac Kinde, Natalie Silliman, Mark Tacey, Hui-Li Wong, Michael Christie, et al. "Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer". *Science Translational Medicine* 8.346 (2016). Developed the algorithm for classifying ctDNA status, analyzed and interpreted data.

[7] **Lu Li**, Yan Wang, YiFan Zhang, Bohao Tang, Ludmila Danilova, Sophie Penisson, Alan Yuille, Linda Chu, and Cristian Tomasetti. "The role of obesity and mutations in cancer". *To be summited*. Obtained all estimates of proportions for risk factors.

## Teaching Experiences

Teaching Assistant at Johns Hopkins University
**2018**: Analysis of Longitudinal Data (140.655), 3rd term

**2017**: Statistical Reasoning in Public Health I & II (140.611 & 140.612), 1st & 2nd term

**2017**: Statistical Reasoning in Public Health I & II (140.611 & 140.612), Summer Institute

**2017**: Multilevel Statistical Models in Public Health (140.656), 4th term

**2017**: Lead TA for Analysis of Longitudinal Data (140.655), 3rd term

**2016**: Statistical Reasoning in Public Health I (140.611), 1st term

**2016**: Analysis of Longitudinal Data (140.608), Summer Institute

**2016**: Statistical Methods in Public Health III & IV (140.623-4), 3rd & 4th term

**2015**: Public Health Biostatistics (AS.280.345), Fall 2015

**2013**: Monte Carlo Methods (EN.550.433), Fall 2013

**2013**: Introduction to Probability (EN.550.420), Spring 2013

## Editorial Activities

Referee for:
PLoS ONE (4)
Journal of Medical Internet Research (2)
JMIR Cancer (1)

## Honors and Awards

**2015**: Best Performance in First Year PhD Comprehensive Exam, Department of Biostatistics, Johns Hopkins University

**2009-2012**: University Scholarship, Wuhan University

## Programming Skills

- Proficient in R, MATLAB, Python, SQL